





# AUTOMATIC AND REMOTE CONTROL

# AUTOMATISME ET TÉLÉCOMMANDE

*Compte-Rendus du  
Deuxième Congrès de la  
Fédération Internationale de la Commande Automatique  
(I.F.A.C.)*

*Bâle, Suisse, 1963*

# REGELUNG UND FERNSTEUERUNG

*Bericht  
über den zweiten Kongress  
des Internationalen Verbandes für Automatische Regelung  
(I.F.A.C.)*

*Basel, Schweiz, 1963*

# AUTOMATIC AND REMOTE CONTROL

*Proceedings of  
the Second Congress of the  
International Federation of Automatic Control  
(I.F.A.C.)*

*Basle, Switzerland, 1963*

*Editor*

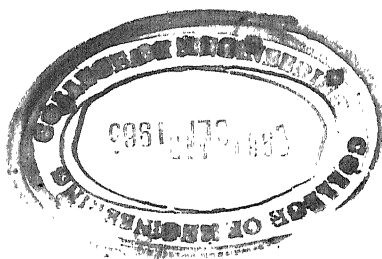
Prof. Dr. Ing. Victor Broïda  
(France)

*Co-editors*

Derek H. Barlow  
(United Kingdom)

Prof. Dr. Ing. Otto Schäfer  
(Germany)

THEORY



BUTTERWORTHS  
LONDON

OLDENBOURG  
MUNICH

1964

BUTTERWORTH & CO. (PUBLISHERS) LTD.

88 Kingsway, London, W.C. 2

Australia: BUTTERWORTH & Co. (AUSTRALIA) LTD.  
SYDNEY: 6-8 O'Connell Street  
MELBOURNE: 473 Bourke Street  
BRISBANE: 240 Queen Street

Canada: BUTTERWORTH & Co. (CANADA) LTD.  
TORONTO: 1367 Danforth Avenue, 6

New Zealand: BUTTERWORTH & Co. (NEW ZEALAND) LTD.  
WELLINGTON: 49-51 Ballance Street  
AUCKLAND: 35 High Street

South Africa: BUTTERWORTH & Co. (SOUTH AFRICA) LTD.  
DURBAN: 33/35 Beach Grove

U. S. A. BUTTERWORTH INC.  
WASHINGTON, D. C.: 7235 Wisconsin Avenue, 14

R. OLDENBOURG VERLAG

Germany: 8 München, Rosenheimer Strasse 145

Austria: Wien III, Neulinggasse 26

©

The several contributors named within  
1964

PGE / N - 14 .

## EDITORIAL NOTE

To surpass success is difficult, but through teamwork and co-operative effort it can be done, as in the case of the Second Congress of the International Federation of Automatic Control.

The first I.F.A.C. Congress was held in Moscow in June-July 1960 and attracted some 1500 participants, and 286 papers which were published in four volumes of some 2000 pages only 15 months later; this First Congress set a high standard for subsequent ones to match.

Yet as judged by its participants, the Second I.F.A.C. Congress held at Basle in August-September 1963 did indeed match these standards. Although the number of papers was only 159, around 1500 participants accompanied by 200 ladies came from 32 different countries to attend these presentations.

Editing these papers into their present form also presented difficulties if the high standard attained with the Proceedings of the First Congress was to be maintained. A lot of these difficulties are, of course, common to all international congresses, but these become aggravated when several languages are used for the original manuscripts of the papers. In the Second Congress, 36 papers were originally submitted in Russian for publication in English, while nine papers were written and published in French. To make the operation truly international, an English Publisher was chosen working in conjunction with a German Printer.

Other difficulties that plague Editors of Congress Proceedings are delays in receiving the manuscripts, rapid translation requirements, and the lack of translators sufficiently expert in the intricacies and detailed knowledge of the field involved. All these problems were present when preparing the preprints of the papers for the Basle Congress. As a result, translations in the preprints were often not of the standard that one would have desired, since the inflexible publication date for the preprints prevented execution of revisions and changes that should have been made.

Even with all the goodwill and co-operation that the Printers and Publishers gave to the project, the 600 mile separation and language difference took their toll. Many typographical errors resulted in the preprints as there was no time for authors to receive and return proofs before the Congress. In spite of conscientious work by both Publisher and Printer after the Congress to check the 1.5 million words in the Proceedings, to improve the translations and to correct the typographical errors, it could be that some traces of these pre-Congress difficulties are still visible. If so, the Editors express their sincerest apologies for any errors that may still be found in these volumes.

Editing of the discussion remarks was greatly eased by the outstanding organization of the Congress by the Swiss Federation of Automatic Control, Professor Ed. Gerecke, Third I.F.A.C. President (1961-1963) and by the Congress Secretary Dr. A. von Schult-hess and the I.F.A.C. Honorary Secretary Dr. G. Ruppel. The main burden in preparing the edited discussion remarks was shouldered by Dipl.-Ing. E. Ruosch of Zurich. Leading an efficient team of scientific secretaries from the Eidgenössische Technische Hochschule of Zurich, the Technische Hochschule of Darmstadt and the University of Cambridge, Dipl.-Ing. Ruosch succeeded in supplying, often less than 48 hours after the presentation of each session, typewritten documents in English of all the discussion remarks on each paper in the session.

In the tedious and painstaking work of editing these volumes, the work of the Co-editors, Otto Schäfer of Aachen and Derek Barlow of London, has been invaluable, especially that of Mr. Barlow. Located in London and hence closest to the Publisher, Mr. Barlow bore the brunt of the detailed editing work, in particular the editing of the discussions on the applications and components papers. Editing of the final form of the Theory papers was done through the co-operation of the U.S.S.R. National Committee on Automatic Control.

Thanks are due also to the Publisher, particularly for the large amount of detailed sub-editing carried out by his co-operative staff, and also the Printer in overcoming the problems of printing scientific works in languages that were not of his own country.

EDITORIAL NOTE

It is hoped that, despite these problems and their effect on the Proceedings, the latter will become a long-living valued contribution in the development of the science of Automatic Control. The excellence of the Congress and the quality of the contributions certainly deserve this recognition.

It has been especially gratifying to the Editor and his colleagues to participate in what has been for them, as well as for all who attended the Congress, one of the most rewarding of human experiences—a fruitful international co-operative effort.

Boulogne sur Seine, France  
June 1964

VICTOR BROÏDA

# CONTENTS

EDITORIAL NOTE . . . . .	V
LIST OF CONTRIBUTORS . . . . .	XXXI

## INTRODUCTORY SPEECHES

ADDRESS OF WELCOME . . . . .	PROFESSOR ED. GERECKE ( <i>President International Federation of Automatic Control</i> )	XXXV
HOW I.F.A.C. WAS FOUNDED . . . . .	H. CHESTNUT ( <i>First President, International Federation of Automatic Control, 1957-59</i> )	XXXVI
THE FIRST I.F.A.C. CONGRESS . . . . .	PROFESSOR A. M. LETOV ( <i>Second President, International Federation of Automatic Control, 1959-61</i> )	XXXVII
THE INFORMATION EVOLUTION AND ITS IMPACT ON AUTOMATIC CONTROL . . . . .	J. L. AUERBACH ( <i>President of the International Federation of Information Processing</i> )	XXXVIII

## THEORY

### NON-LINEAR SYSTEM THEORY

Statistical Methods in Automatic Control — A survey . . . . .	V. S. Pugachev	1
<b>Describing Function Technique</b>		
A New Method to Derive the Describing Function of Certain Non-linear Transfer Systems . . . . .	R. Lauber	14
The Describing Function Method Applied for the Investigation of Parametric Excited Oscillations . . . . .	A. Leonhard	21
On the Inverse Describing Function Problem . . . . .	J. E. Gibson and E. S. di Tada	29
Relative Stability of Oscillations in Non-linear Control Systems . . . . .	Z. Bonenn	35
<b>Prediction Systems</b>		
Predictive Control of an On-Off System with Two Control Variables . . . . .	A. J. Adey, J. F. Coales and J. A. Stiles	41
A Method of Prediction for Non-stationary Processes and Its Application to the Problem of Load Estimation . . . . .	E. D. Farmer	47
On the Design of Predictor Control Systems . . . . .	S. Horing	55
A Method of Optimal Control Prediction . . . . .	F. B. Gul'ko and B. Ya. Kogan	63
<b>Non-linear Systems with Random Parameters</b>		
Optimalization of Non-linear Random Control Processes . . . . .	R. Kulikowski	69
Non-linear Programming in the Investigation of Optimal Automatic Control Systems . . . . .	N. I. Andreev	76
Numerical Analysis of Non-linear Control Systems Using the Fokker-Planck-Kolmogorov Equation . . . . .	K. J. Merklinger	81
Volterra Series Representation of Time-varying Non-linear Systems . . . . .	R. H. Flake	91
<b>Non-linear Stochastic Systems</b>		
The Applicability of Quasi-linear Methods to Non-linear Feedback Systems with Random Inputs . . . . .	H. W. Smith	100
A Digital Procedure for the Study of Non-linear Systems for Random Processes . . . . .	T. Prasad and V. P. Sinha	108
Time-optimal Systems with Random Noise Disturbances . . . . .	V. N. Novoseltsev	114

## DISCRETE SYSTEMS

## Discrete Systems

Quasi-invariant Hybrid Multi-parameter Control Loops . . . . .	<i>V. Strejc</i>	122
Optimum Control of Discrete-time Dynamic Processes . . . . .	<i>B. Friedland</i>	128
Theory of Minimum Time Discrete Regulators . . . . .	<i>C. A. Desoer, E. Polak and J. Wing</i>	135

## Sampled-data Systems

On the Roots of a Real Polynomial Inside the Unit Circle and a Stability Criterion for Linear Discrete Systems . . . . .	<i>E. I. Jury</i>	142
Analytical Approaches to Non-linear Sampled-data Control Systems . . . . .	<i>B. Kondo and S. Iwai</i>	154
Continuous Compensation of Feedback Sampled-data Linear Control Systems . . . . .	<i>B. M. Brown</i>	165
Fundamentals of the Theory of Non-linear Sampled-data Control Systems . . . . .	<i>Ya. Z. Tsypkin</i>	172
Synthesis of Optimum Sampled-data Systems . . . . .	<i>L. N. Volgin</i>	181
Integral Pulse Frequency Modulated Control Systems . . . . .	<i>C. C. Li and R. W. Jones</i>	186
Combination of Finite Settling Time and Minimum Integral of Squared Error in Digital Control Systems . . . . .	<i>V. Peterka</i>	196

## Relay Systems

Oscillations sous-harmoniques dans un asservissement par plus-ou-moins . . . . .	<i>J. C. Gille, S. Węgrzyn and J.-G. Paquet</i>	204
Synthesis of Control Systems Operating Linearly for Small Signals and Approximately 'Bang-Bang' for Large Signals . . . . .	<i>E. V. Persson</i>	210
Dual Input Systems with a Saturation Constraint . . . . .	<i>R. S. Gaylord</i>	219

## Theory of Finite Automata

Signalling and Prediction of Failures in Discrete Control Devices with Structural Redundancy . . . . .	<i>M. A. Gavrilo</i>	228
Axiomatization of the Theory of Simplification of Combinational Automata . . . . .	<i>Gr. C. Moisil</i>	234
Adaptive Control for a System with a Finite Number of States . . . . .	<i>S. Paszkowski</i>	241

## OPTIMAL SYSTEMS

Synthesis of Optimal Regulators—A survey . . . . .	<i>A. M. Letov</i>	246
--	--------------------	-----

## Optimal Systems

Approximation Methods in Optimal and Adaptive Control . . . . .	<i>J. H. Westcott, J. J. Florentin and J. D. Pearson</i>	263
An Optimal Guidance Approximation for Quasi-circular Orbital Rendezvous . . . . .	<i>H. J. Kelly and J. C. Dunn</i>	274
On Synthesizing Optimal Controls . . . . .	<i>L. W. Neustadt and B. H. Paiewonsky</i>	283
An Application of Optimal Control to Midcourse Guidance . . . . .	<i>J. S. Meditch and L. W. Neustadt</i>	292
On Necessary and Sufficient Conditions for Time-optimal Control of Second-order Non-linear Systems . . . . .	<i>E. B. Lee and L. Markus</i>	300
On Optimal Control of Systems with Multi-norm Constraints . . . . .	<i>P. E. Sarachik and G. M. Kranc</i>	306
The Approximate Calculation of a Class of Automatic Systems with Forced Parameter Optimization . . . . .	<i>Yu. I. Alimov</i>	315

## Optimal Systems with Distributed Parameters

Some Considerations on Optimized Integrated Control . . . . .	<i>R. Marcacci</i>	324
Optimal Processes in Systems with Time Lag . . . . .	<i>N. N. Krasovskii</i>	327
Optimal Control of Systems with Distributed Parameters . . . . .	<i>A. G. Butkovskii</i>	333

## Synthesis of Optimum Systems

Solution of Optimum Control Problems by Using Pontryagin's Maximum Principle . . . . .	<i>Y. Sakawa and C. Hayashi</i>	339
Analysis and Synthesis of Time-optimal Control Systems . . . . .	<i>Sun Jian and Hang King-ching</i>	347
Most Recent Development of Dynamic Programming Techniques and Their Application to Optimal Systems Design . . . . .	<i>R. L. Stratonovich</i>	352
A Modified Maximum Principle for Optimum Control of a System with Bounded Phase Space Coordinates . . . . .	<i>S. S. L. Chang</i>	358
Optimum and Quasi-optimum Control of Third- and Fourth-order Systems . . . . .	<i>I. Flügge-Lotz and H. A. Titus, Jr.</i>	363

## Optimal Programming

Programme Control and the Theory of Optimal Systems . . . . .	<i>Ye. A. Barbashin</i>	371
The Realization of Optimal Programmes in Control Systems . . . . .	<i>G. S. Pospelov</i>	377
Some Bounds on Quantization Errors in Dynamic Programming Computations . . . . .	<i>J. J. G. Guignabodet</i>	383



# CONTENTS

## THEORY OF SELF-ADJUSTING SYSTEMS

Adaptive Control—A survey . . . . .	<i>J. Truxal</i>	386
Learning Systems—A survey . . . . .	<i>G. Pask</i>	393

### Invariancy Problems

Invariance of Sampled-data and Adaptive Sampled-data Systems . . . . .	<i>V. M. Kuntsevich and Yu. V. Krementulo</i>	412
Optimization and Invariance in Control Systems with Constant and Variable Structure . . . . .	<i>B. N. Petrov, G. M. Ulanov and S. V. Emelyanov</i>	421
Synthesis of Systems with Fixed Characteristics of Equivalent Self-adjusting Systems . . . . .	<i>M. V. Meerov</i>	430

### Self-adapting Systems

A Comparison of the Measuring Time in Self-adjusting Control Systems . . . . .	<i>F. Mesch</i>	439
On Self-adjusting Control Systems without Test Disturbance Signals . . . . .	<i>E. P. Popov, G. M. Loskutov and R. M. Yusupov</i>	446
On the Dynamics of Some Learning and Self-learning Processes . . . . .	<i>V. K. Chichinadze</i>	453
On the Searching of Extrema of Functions in Automatic Control Systems . . . . .	<i>A. A. Voronov and M. B. Ignatjev</i>	459
Dominant Operators' Approach to the Theory of Adaptive Control Systems . . . . .	<i>A. Straszak</i>	465
General Stability Analysis of Sinusoidal Perturbation Extrema Searching Adaptive Systems . . . . .	<i>V. W. Eveleigh</i>	472
The Realization of a Self-adapting Control Programme in a System with a Digital Calculating Computer . . . . .	<i>P. F. Klubnikin</i>	481

### Learning Systems

A Pattern Recognizing Adaptive Controller . . . . .	<i>W. K. Taylor</i>	488
STELLA: A Scheme for a Learning Machine . . . . .	<i>J. H. Andrae</i>	497
Automatic Control Learning Systems (in the Light of Experiments on Teaching the Systems to Pattern Recognition) . . . . .	<i>M. A. Aizerman</i>	503

### Hill-climbing Technique

On the Theory of Self-tuning Systems with a Search for Gradient by the Method of Auxiliary Operator . . . . .	<i>I. E. Kazakov and L. G. Evlanov</i>	510
Problems of Continuous Systems Theory of Extremal Control of Industrial Processes . . . . .	<i>A. A. Krasovski</i>	519
Principle and Application of an Extremal Computer . . . . .	<i>R. Perret and R. Rouxel</i>	527
Dual Control Theory Problems . . . . .	<i>A. A. Feldbaum</i>	541
The Application of Random Test Signals in Process Optimization . . . . .	<i>P. M. E. M. van der Grinten</i>	551

## TECHNIQUES FOR SYSTEM STABILITY ASSESSMENT

### Liapunov's Stability Problem

Eventual Stability . . . . .	<i>J. P. LaSalle and R. J. Rath</i>	556
Non-linear Stability Analysis for Restricted Non-linearities Using the Second Method of Liapunov . . . . .	<i>H. Nour-Eldin</i>	561
The Use of the Technique of Linear Bounds for Applying the Direct Method of Liapunov to a Class of Non-linear and Time-varying Systems . . . . .	<i>R. A. Nesbit</i>	568
On the Estimation of the Decaying Time . . . . .	<i>L. Hwang</i>	576
New Methods for Constructing Liapunov Functions for Time-invariant Control Systems . . . . .	<i>G. P. Szegö</i>	584
A Method of Investigating Stability . . . . .	<i>H. H. Rosenbrock</i>	590

# CONTENTS

## SYSTEM DYNAMICS AND OTHER PROBLEMS

Process Dynamics and its Application to Industrial Process Design and Process Control—A survey . . . . .	<i>T. J. Williams</i>	595
--	-----------------------	-----

### System Dynamics

Determination of System Dynamics by Use of Adjustable Models . . . . .	<i>E. Blandhol and J. G. Balchen</i>	602
Les erreurs systématiques et aléatoires dans la détermination expérimentale des fonctions de transfert . . . . .	<i>J. Loeb</i>	614
The Applicability of the Relay Correlator and the Polarity Coincidence Correlator in Automatic Control . . . . .	<i>B. P. Th. Veltman and A. van den Bos</i>	620
Notes sur une fonction aléatoire d'aspect physique . . . . .	<i>M. J. Pélégryn</i>	628
An Uncertainty Relation for Linear Mathematical Models . . . . .	<i>B. Qvarnström</i>	634

### General Problems

Axiomatic Foundation of the Theory of Control Systems . . . . .	<i>E. Roxin</i>	640
The Inverse Problem of Integral Square Estimation of Transient Responses . . . . .	<i>W. Jarominek</i>	645
On Systems with Automatic Control of Configuration . . . . .	<i>J. Beneš</i>	656
Approximation of Industrial Control Systems for Optimized Non-linear Control with Restricted Rate of Correction Using Conventional Controllers . . . . .	<i>E. Pavlik</i>	664
Nouvelle procédure d'optimisation statistique fondée sur la transformation $V = (Z - 1)/(Z + 1)$ . . . . .	<i>P. M. Lefèvre</i>	671
The Control of Two Output-dependant Processes . . . . .	<i>C. N. Kerr and G. D. S. MacLellan</i>	682

## REPORTS

<i>Academician B. N. Petrov</i> (Chairman of the I.F.A.C. Technical Committee on Theory) . . . . .	693
<i>Professor J. H. Westcott</i> (Vice-Chairman of the I.F.A.C. Technical Committee on Theory) . . . . .	694

## List of contents of the volume on Applications and Components

# APPLICATIONS

## THE ELECTRICAL UTILITY FIELD

Applications of Automatic Control in Electrical Utility Systems—A survey . . . . . *B. Favez*

### Electric Utility

Working Out a Method for the Cybernetic Control of Integrated Electric Power Systems . . . *V. A. Venikov and L. V. Tsukernik*  
Some Recent Results in the Computer Control of Energy Systems . . . . . *T. Vámos, S. Benedikt and M. Uzsoki*  
Optimal Control of Thermal-hydro System Operation . . . . . *L. K. Kirchmayer and R. J. Ringlee*  
Automatic Systems with Learning Elements . . . . . *G. K. Krug and A. V. Netushil*  
Principes et réalisations des automatismes liés à la manutention de combustible d'un réacteur nucléaire . . *P. Turpin and J. Thilliez*  
A Study of the Dynamics of Steam Voids in Boiling Water Nuclear Reactors . . . . . *P. K. M'Pherson and M. Muscettola*

### Hydro-system Control

On the Optimal Control of Hydro-electric Power Systems . . . . . *H. Ruge*  
Exposé d'une méthode d'élaboration de graphiques exprimant les conditions de stabilité du réglage d'un groupe hydro-électrique  
*A. Tschumy*  
Digital Investigation of Multi-machine Power Systems . . . . . *H. Glavitsch*

### Electric Utility-Machine Control

Optimizing Control of Water Turbine Governors Considering the Non-linearity of Servomotor Speed . . . . . *T. Stein*  
Control Equations of a Hydro-electric Plant with Fixed Reference Values . . . . . *L. Borel*

## THE STEEL INDUSTRY

Application of Automation and Automatic Techniques to Metal Rolling and Processing—A survey . . . . . *W. E. Miller*  
Achievements in the Automation of Ferrous Metallurgy—A survey . . . . . *A. Ya. Lerner*

### Steel Industry

Control for the Sintering Mixture Preparation . . . . . *G. DeGregorio, G. Litigio and G. Sironi*  
Dynamic Planning for an Open-hearth Steel-making Plant . . . . . *M. Korobko and Yu. Samoilenko*  
Automation of Heavy Forging . . . . . *J. G. Wistreich and A. Tomlinson*  
Automation in a Steel Works with Special Reference to the Use of Digital Computers for Production Scheduling and Information Transmission . . . . . *S. E. Hersom and R. G. Massey*  
On-line Computer Control of a Hot Strip Finishing Mill for Steel . . . . . *R. G. Beadle*  
Optimum Control for Continuous Processes . . . . . *A. Ya. Lerner*  
A Digital Optimal System of Programmed Control and Its Application to the Screw-down Mechanism of a Blooming Mill  
*S. M. Domanitsky, V. V. Imedadze and Sh. A. Tsintsadze*  
Computer Control of the Continuous Annealing Process . . . . . *J. T. Bradford, Jr.*

## THE CHEMICAL AND OIL INDUSTRIES

Application of Automatic Control in the Chemical and Oil Industries—A survey

*H. W. Slotboom, J. J. de Jong, J. A. Landstra, J. E. Rijnsdorp and A. C. Timmens*

## Chemical and Oil Industries

- Dynamic Characteristics of Binary Distillation Column . . . . . *K. Izawa and T. Morinaga*  
 Approximation Models for the Dynamic Response of Large Distillation Columns . . . . . *J. S. Moczek, R. E. Otto and T. J. Williams*  
 A Study on the Dynamic Behaviour of a Catalytic Cracker Power-recovery System by Means of an Analogue Computer  
     *C. A. J. M. van der Heyden and A. G. van Nes*  
 Statistical Analysis of a Novel Fluid Flow Control System . . . . . *R. C. Booton, Jr. and W. E. Sollecito*  
 Effects of Fluid Mixing and Its Expressions on Dynamics of Mass Transfer Process . . . . . *T. Takamatsu and E. Nakanishi*  
 Experimental Study of the Dynamic Behaviour of a Heat Exchanger and of a Mixing Process . . . . . *L. Delvaux*  
 Study of Industrial Production of Polyethelene Under High Pressures, and of the Automatic Control of the Process . . . . . *B. V. Volter*  
 A Study of the Dynamic and Static Characteristics of the Process of Fractional Distillation . . . . . *I. V. Anisimov*  
 Controllability and Allowable Compressor Capacity of a Flare Gas Recovery System . . . . . *F. J. Kylstra*  
 Analysis and Design of a Parameter-perturbation Adaptive System for Application to Process Control . . . . . *T. Isobe and T. Totani*  
 The Dynamic Properties of Rectification Stations with Plate Columns . . . . . *J. Závorka*  
 Une réalisation originale dans une raffinerie de pétrole: le chargement automatique des wagons citernes . . . . . *F. X. Montjean*

## AUTOMATION IN INDUSTRIAL PROCESSES

## Computer On- and Off-line

- Le traitement du problème d'optimalisation par A. 110 . . . . . *E. Honoré*  
 Automation of a Portland Cement Plant Using a Digital Control Computer . . . . . *R. A. Phillips*  
 On the Stability and Design of Dither Adaptive Systems . . . . . *R. K. Smyth and N. E. Nahi*

## Steam-system Control

- Predetermination of Control Results for Reheaters in Steam Generators . . . . . *W. Kindermann*  
 An Optimizing Control of Boiler Efficiency . . . . . *S. Fujii and N. Kanda*  
 Simulator for Steam Turbines with Reheat or Automatically Controlled Extraction . . . . . *H.-J. Ehling*  
 An Electro-hydraulic Control System for Reheat Turbines . . . . . *M. A. Eggenberger and P. H. Troutman*

## Automotive Industry

- Une application industrielle d'un calculateur intérieur à un circuit de commande: l'équipement électronique de commande d'une machine à équilibrer les vilebrequins . . . . . *J. Csech*  
 Design Analysis of an Automotive Speed Control System . . . . . *W. H. Holl*

## COMBINED MAN-MACHINE SYSTEMS

## Man and Machine

- Modèles continus et échantillonnés de l'opérateur humain placé dans une boucle de commande . . . . . *P. Naslin and J.-C. Raoult*  
 Discrete Models of the Human Operator in a Control System . . . . . *G. E. Bekey*  
 Dynamic Analysis and Simulation of Management Control Functions . . . . . *R. B. Wilcox*  
 Outline of a Control Theory of Prosthetics . . . . . *R. Tomovic*  
 A Sampled-data Model for Eye Tracking Movements . . . . . *L. R. Young*

## APPLICATION TECHNIQUES

## Methods

- Self-adaptive Method for Accommodating Large Variations of Plant Gain in Control Systems . . . . . *R. J. Kochenburger*  
 Hydraulic Line Dynamics . . . . . *R. Oldenburger and R. E. Goodson*  
 The Optimization of Computer-controlled Systems Using Partial Knowledge of the Output State . . . . . *B. G. Anderson*

CONTENTS

AUTOMATIC CONTROL OF AEROSPACE SYSTEMS

Aerospace Systems

The Design Study of a Pressure Control System for a 5 ft. by 5 ft. Blowdown Wind Tunnel . . . . .	<i>J. A. Tanner and W. Dietiker</i>
Missile Environment Simulation for Rocket Engine Test Facility . . . . .	<i>G. J. Fiedler and J. J. Landy</i>
A Longitudinal Guidance System for Aircraft Landing During Flare-out . . . . .	<i>F. J. Ellert and C. W. Merriam III</i>
Automatic Control of a Large Steerable Aerial for Satellite Communications . . . . .	<i>F. J. D. Taylor</i>
Design Study of a Control System for a 210 ft. Radio Telescope . . . . .	<i>R. G. Wheeler</i>
Dynamical Model for Fine Pointing Attitude Control of the Orbiting Astronomical Observatory . . . . .	<i>R. E. Roberson</i>
Basic Response Relations for Satellite Attitude Control Using Gyros . . . . .	<i>R. H. Cannon, Jr.</i>

REPORT

<i>W. E. Miller</i> (Chairman of the I.F.A.C. Technical Committee on Applications) . . . . .	
--	--

# COMPONENTS

Control Components—New Design Principles and Control Devices—A survey . . . . . *J. L. Shearer and 16 co-authors*

## MECHANICAL, HYDRAULIC AND PNEUMATIC DEVICES

### Mechanical, Hydraulic and Pneumatic Devices

A Hydraulic Torque Amplifier . . . . . *Y. Oshima and K. Araki*  
 A Rotary-drive Vibratory-output Gyroscopic Instrument . . . . . *G. C. Newton*  
 Some Problems of the Dynamics of a Hydraulic Throttle-control Servo-mechanism with an Inertial Load . . . . . *V. Khokhlov*  
 Realization of Sequential Machines by Means of Pneumatic Automation . . . . . *A. A. Tal*

## ELECTROMECHANICAL DEVICES AND MAGNETIC AMPLIFIERS

### Electromechanical Devices and Magnetic Amplifiers

Two-positional Functional Frequency Device for Automatic Regulation . . . . . *I. A. Maslaroff*  
 The Problems, Operation and Calculation of a New Component to be Applied in Certain Control Circuits . . . . . *O. Benedikt*  
 The Control of a Linear Electromagnetic Oscillating Mechanism . . . . . *J. C. West and B. V. Jayawant*  
 Step Motors with an Active Rotor . . . . . *Yu. K. Vasilev, Yu. A. Prokofiev and G. Ya. Wainberger*

## ELECTRONIC COMPONENTS

### Electronic Components

Some New Control Circuits Using Four-layer  $p-n-p-n$  Semiconductor Triodes . . . . . *J. Š. Haškovec and A. Klímek*  
 Turn-off Silicon Controlled Rectifiers . . . . . *H. F. Storm*  
 Ceramic Memories in Extreme Environments . . . . . *A. B. Kaufman*

## DIGITAL DEVICES

### Digital Devices

The Application of Digital Differential Analysers in Control Loops . . . . . *H. Rechberger*  
 Advantages and Possibilities of Digital Speed Control . . . . . *W. Fritzsche*  
 Digital Controllers . . . . . *T. M. Aleksandridi, S. N. Diligensky and H. K. Krug*

## PROCESS INSTRUMENTATION

### Various Components

A Universal Statistical Analyser . . . . . *J. Krýže*  
 The Behaviour of Adaptive Controllers . . . . . *J. L. Douce*  
 The Design Principles and Circuit of a Multi-channel Correlator—a Specialized Analogue Computer for the Statistical Treatment of Random Time Series in Industrial Control Systems . . . . . *A. S. Uskov and Yu. M. Orlov*

## COMPONENT RELIABILITY

The Reliability of Components—A survey . . . . . *G. Glinski, B. S. Sotskov and H. Weissmann*

### Reliability

New Servovalves for Redundant Electrohydraulic Control . . . . . *K. D. Garnjost and W. J. Thayer*  
 The Reliability of Electronic Components . . . . . *G. W. A. Dummer*  
 Reliability Problems of Electromechanical Elements . . . . . *B. S. Sotskov, I. E. Dekabrun and L. S. Krivorotova*  
 A Study of Servomechanism Reliability in Nuclear Reactor and Plant Control Systems . . . . . *L. A. J. Lawrence and R. J. Scotcher*

## REPORTS

*Gy. Boromisza* (Chairman of the I.F.A.C. Technical Committee On Components) . . . . .  
*Y. Oshima* (Vice-Chairman of the I.F.A.C. Technical Committee On Components) . . . . .

# TABLE DES MATIÈRES

AVANT-PROPOS . . . . .	V
LISTE DES AUTEURS . . . . .	XXXI

## EXPOSÉS INTRODUCTIFS

ALLOCUTION DE BIENVENUE . . . . .	PROFESSEUR ED. GERECKE (Président de l'I.F.A.C.)	XXXV
COMMENT L'I.F.A.C. A-T-ELLE ÉTÉ FONDÉE . . . . .	H. CHESTNUT (Premier Président de l'I.F.A.C. 1957-1959)	XXXVI
LE PREMIER CONGRÈS DE L'I.F.A.C. . . . .	PROFESSEUR A. M. LETOV (Deuxième Président de l'I.F.A.C. 1959-1961)	XXXVII
L'ÉVOLUTION DE MÉTHODES D'INFORMATION ET SA RÉPERCUSSION SUR LES MÉTHODES DE COMMANDE AUTOMATIQUE . . . . .	J. L. AUERBACH (Président de l'I.F.I.P.)	XXXVII

## THÉORIE

### LA THÉORIE DES SYSTÈMES NON-LINÉAIRES

Les méthodes statistiques dans la commande automatique — Une synthèse . . . . .	V. S. Pugachev	1
<b>Technique de la fonction descriptive</b>		
Une nouvelle méthode de détermination de la fonction descriptive de certains systèmes de transfert non-linéaires . . . . .	R. Lauber	14
Application de la méthode de la fonction descriptive à l'étude d'oscillations forcées paramétriques . . . . .	A. Leonhard	21
Sur le problème de la fonction descriptive inverse . . . . .	J. E. Gibson et E. S. di Tada	29
La stabilité relative d'oscillations de systèmes non-linéaires . . . . .	Z. Bonem	35
<b>Systèmes d'anticipation</b>		
La commande anticipante d'un système à deux variables par tout ou rien . . . . .	A. J. Adey, J. F. Coales et J. A. Stiles	41
Une méthode d'anticipation de processus non-stationnaires et son application au problème de l'analyse des variations de charge . . . . .	E. D. Farmer	47
Sur le calcul des systèmes de commande à anticipation . . . . .	S. Horing	55
Méthode d'optimisation avec anticipation . . . . .	F. B. Gul'ko et B. Ya. Kogan	63
<b>Systèmes non-linéaires à paramètres aléatoires</b>		
Optimisation de processus de commande non-linéaires aléatoires . . . . .	R. Kulikowski	69
La programmation non-linéaire dans l'étude des systèmes de commande optimaux . . . . .	N. I. Andreev	76
Analyse numérique des systèmes de commande non-linéaires par l'emploi de l'équation de Fokker-Planck-Kolmogorov . . . . .	K. J. Merklinger	81
La représentation au moyen de séries de Volterra de systèmes non-linéaires variant avec le temps . . . . .	R. H. Flake	91
<b>Systèmes aléatoires non-linéaires</b>		
L'application de méthodes quasi-linéaires aux systèmes asservis non-linéaires à grandeurs d'entrée aléatoires . . . . .	H. W. Smith	100
Une méthode numérique d'étude de systèmes non-linéaires pour processus aléatoires . . . . .	T. Prasad et V. P. Sinha	108
Systèmes d'optimisation selon le temps avec perturbations de bruits aléatoire . . . . .	V. N. Novoseltsev	114

## SYSTÈMES DISCRETS

Systèmes d'asservissement hybrides quasi-invariants à paramètres multiples . . . . .	<i>V. Strejc</i>	122
Optimalisation de processus dynamiques à temps discret . . . . .	<i>B. Friedland</i>	128
Théorie des régulateurs discrets à minimum de temps . . . . .	<i>C. A. Desoer, E. Polak et J. Wing</i>	135

## Systèmes échantillonnés

Sur les racines d'un polynôme réel à l'intérieur du cercle unitaire et un critère de stabilité pour systèmes discrets linéaires . . . . .	<i>E. I. Jury</i>	142
Manières analytiques d'envisager les systèmes de commande échantillonnés non-linéaires . . . . .	<i>B. Kondo et S. Iwai</i>	154
Compensation continue des systèmes de réglage échantillonnés linéaires . . . . .	<i>B. M. Brown</i>	165
Les fondements de la théorie des systèmes de commande échantillonnés non-linéaires . . . . .	<i>Ya. Z. Tsypkin</i>	172
La synthèse des systèmes échantillonnés optimaux . . . . .	<i>L. N. Volgin</i>	181
Systèmes de commande à modulation par la fréquence d'impulsions intégrale . . . . .	<i>C. C. Li et R. W. Jones</i>	186
Combinaison du temps d'établissement fini et du minimum du carré moyen de l'erreur de réglage dans les systèmes de commande numériques . . . . .	<i>V. Peterka</i>	196

## Systèmes à relais

Oscillations sous-harmoniques des asservissements par plus-ou-moins . . . . .	<i>J. C. Gille, S. Wegrzyn et J.-G. Paquet</i>	204
Synthèse des systèmes de commande fonctionnant d'une manière linéaire pour les faibles signaux et d'une manière discontinue pour les signaux importants . . . . .	<i>E. V. Persson</i>	210
Systèmes à deux entrées avec une contrainte de saturation . . . . .	<i>R. S. Gaylord</i>	219

## Théorie des automates à positions multiples

Signalisation et prévision de pannes dans les dispositifs de commande discrète à surabondance de moyens inhérente . . . . .	<i>M. A. Gavrilo</i>	228
Axiomatisation de la théorie de simplification des automates combinatoires . . . . .	<i>Gr. C. Moisil</i>	234
Commande adaptative d'un système présentant un nombre fini d'états . . . . .	<i>S. Paszkowski</i>	241

## THÉORIE DES SYSTÈMES OPTIMAUX

Synthèse des régulateurs optimaux — Une synthèse . . . . .	<i>A. M. Letov</i>	246
--	--------------------	-----

## Systèmes optimaux

Méthodes d'approximation dans la commande optimale et adaptative . . . . .	<i>J. H. Westcott, J. J. Florentin et J. D. Pearson</i>	263
Une approximation optimale de guidage pour le rendez-vous de deux satellites sur une orbite quasi-circulaire . . . . .	<i>H. J. Kelly et J. C. Dunn</i>	274
Sur la synthèse des commandes optimales . . . . .	<i>L. W. Neustadt et B. H. Paiewonsky</i>	283
Une application de la commande optimale au guidage d'engins à mi-course . . . . .	<i>J. S. Meditch et L. W. Neustadt</i>	292
Sur les conditions nécessaires et suffisantes pour l'optimalisation selon le temps de systèmes non-linéaires du second ordre . . . . .	<i>E. B. Lee et L. Markus</i>	300
L'optimalisation des systèmes de commande avec contraintes à normes multiples . . . . .	<i>P. E. Sarachik et G. M. Kranc</i>	306
Le calcul approximatif d'une catégorie de systèmes automatiques avec optimalisation forcée des paramètres . . . . .	<i>Yu. I. Alimov</i>	315

## Systèmes optimaux à paramètres repartis

Quelques considérations sur la commande optimisée intégrée . . . . .	<i>R. Marcelli</i>	324
Processus optimaux dans les systèmes avec constantes de temps . . . . .	<i>N. N. Krasovskii</i>	327
Commande optimale des systèmes à paramètres repartis . . . . .	<i>A. G. Butkovskii</i>	333

## Synthèse des systèmes optimaux

Solution des problèmes d'optimalisation en utilisant le principe du maximum de Pontryagin . . . . .	<i>Y. Sakawa et C. Hayashi</i>	339
Analyse et synthèse des systèmes d'optimalisation selon le temps . . . . .	<i>Sun Jian et Hang King-ching</i>	347
Développement des techniques de programmation dynamique et leurs applications à l'étude des systèmes optimaux . . . . .	<i>R. L. Stratonovich</i>	352
Un principe du maximum modifié pour l'optimalisation d'un système dont les cordonnées dans l'espace de phase sont limitées . . . . .	<i>S. S. L. Chang</i>	358
Commande optimale et quasi-optimale de systèmes du troisième et du quatrième ordres . . . . .	<i>I. Flügge-Lotz et H. A. Titus, Jr.</i>	363



**Programmation optimale**

Commande programmée et théorie des systèmes optimaux . . . . .	<i>Ye. A. Barbashin</i>	371
La réalisation de programmes optimaux dans les systèmes de commande . . . . .	<i>G. S. Pospelov</i>	377
Quelques limites des erreurs de quantification dans les calculs numériques de la programmation dynamique . . . . .	<i>J. J. Guignabodet</i>	383

**THÉORIE DES SYSTÈMES A AUTO-RÉGLAGE**

Systèmes adaptatifs et à auto-optimalisation — Une synthèse . . . . .	<i>J. Truxal</i>	386
Systèmes à apprentissage — Une synthèse . . . . .	<i>G. Pask</i>	393

**Problèmes d'invariance**

L'invariance des systèmes échantillonnés et des systèmes adaptatifs échantillonnés . . . . .	<i>V. M. Kuntsevich et Yu. V. Krementulo</i>	412
L'optimalisation et l'invariance des systèmes de commande à structures constante et variable . . . . .	<i>B. N. Petrov, G. M. Ulanov et S. V. Emelyanov</i>	421
Synthèse de systèmes avec les caractéristiques fixes des systèmes à auto-réglage équivalents . . . . .	<i>M. V. Meyerov</i>	430

**Systèmes auto-adaptatifs**

Une comparaison du temps de mesure dans les systèmes de commande à auto-réglage . . . . .	<i>F. Mesch</i>	439
Systèmes de commande adaptative sans signaux d'entrée explorateurs . . . . .	<i>E. P. Popov, G. M. Loskutov et R. M. Yusupov</i>	446
Une contribution à l'utilisation des systèmes à auto-réglage pour la synthèse mécanique des systèmes de commande . . . . .	<i>V. K. Chichinadze</i>	453
Sur la recherche des valeurs extrémales de fonctions dans les systèmes de commande automatique . . . . .	<i>A. A. Voronov et M. B. Ignatjev</i>	459
Méthode de la théorie des systèmes de commande adaptative utilisant les opérateurs dominants . . . . .	<i>A. Straszak</i>	465
Analyse générale de la stabilité des systèmes adaptatifs avec recherche des valeurs extrémales au moyen de perturbations sinusoïdales . . . . .	<i>V. W. Eveleigh</i>	472
La réalisation d'un programme de commande auto-adaptatif dans un système au moyen d'un calculateur numérique . . . . .	<i>P. F. Klubnikin</i>	481

**Systèmes à apprentissage**

Un régulateur adaptatif identifiant les espèces . . . . .	<i>W. K. Taylor</i>	488
STELLA: le schéma d'une machine à apprentissage . . . . .	<i>J. H. Andreae</i>	497
Systèmes de commande automatique à apprentissage (à la lumière des expériences rendant les systèmes aptes à identifier les espèces) . . . . .	<i>M. A. Aizerman</i>	503

**Techniques de recherche de valeurs extrémales successives**

Sur la théorie des systèmes à accord automatique avec une recherche du gradient par la méthode de l'opérateur auxiliaire . . . . .	<i>L. E. Kazakov et L. G. Evlanov</i>	510
Problèmes de la théorie des systèmes continus de commande extrême de processus industriels . . . . .	<i>A. A. Krasovski</i>	519
Principe et application d'un calculateur extrême . . . . .	<i>R. Perret et R. Rouxel</i>	527
Problèmes de la théorie de la commande double . . . . .	<i>A. A. Feldbaum</i>	541
Application des signaux explorateurs aléatoires à l'optimalisation de processus . . . . .	<i>P. M. E. M. van der Grinten</i>	551

**TECHNIQUES D'ÉVALUATION DE LA STABILITÉ DE SYSTÈMES****Problème de stabilité de Liapunov**

Un nouveau concept de stabilité . . . . .	<i>J. P. LaSalle et R. J. Rath</i>	556
Analyse non-linéaire de la stabilité des non-linéarités restreintes en utilisant la seconde méthode de Liapunov . . . . .	<i>H. Nour-Eldin</i>	561
L'utilisation de la technique des contraintes linéaires pour l'application de la méthode directe de Liapunov à une catégorie de systèmes non-linéaires et de systèmes variant en fonction du temps . . . . .	<i>R. A. Nesbit</i>	568
Estimation du temps d'amortissement . . . . .	<i>L. Hwang</i>	576
Nouvelles méthodes pour la détermination des fonctions de Liapunov pour des systèmes de commande invariants dans le temps . . . . .	<i>G. P. Szegö</i>	584
Une méthode d'investigation de la stabilité . . . . .	<i>H. H. Rosenbrock</i>	590

## DYNAMIQUE DES SYSTÈMES ET AUTRES PROBLÈMES

Dynamique des processus et son application à l'étude et à la commande des processus industriels — Une synthèse . . . . .	<i>T. J. Williams</i>	595
--	-----------------------	-----

**Dynamique des systèmes**

Détermination de la dynamique des systèmes en utilisant des modèles réglables . . . . .	<i>E. Blandhol et J. G. Balchen</i>	602
Les erreurs systématiques et aléatoires dans la détermination expérimentale des fonctions de transfert . . . . .	<i>J. Loeb</i>	614
L'applicabilité du corrélateur à relais et du corrélateur à coïncidence de polarités dans la commande automatique . . . . .	<i>B. P. Th. Veltman et A. van den Bos</i>	620
Notes sur une fonction aléatoire d'aspect physique . . . . .	<i>M. J. Pélégryn</i>	628
Une relation d'incertitude pour les modèles mathématiques linéaires . . . . .	<i>B. Qvarnström</i>	634

**Problèmes généraux**

Les fondements axiomatiques de la théorie des systèmes de commande . . . . .	<i>E. Roxin</i>	640
Le problème inverse de l'estimation quadratique moyenne des réponses transitoires . . . . .	<i>W. Jarominek</i>	645
Sur les systèmes à commande automatique de la configuration . . . . .	<i>J. Beneš</i>	656
Approximation des systèmes de commande automatique industrielle pour l'optimisation non-linéaire à vitesse de correction limitée utilisant des régulateurs conventionnels . . . . .	<i>E. Pavlik</i>	664
Nouvelle procédure d'optimisation statistique fondée sur la transformation $V = (Z - 1)/(Z + 1)$ . . . . .	<i>P. M. Lefèvre</i>	671
La commande de processus dépendant de deux grandeurs de sortie . . . . .	<i>C. N. Kerr et G. D. S. MacLellan</i>	682

## RAPPORTS

<i>Académicien B. N. Petrov</i> (Président du Comité de Théorie) . . . . .	693
<i>Professeur J. H. Westcott</i> (Vice-Président du Comité de Théorie) . . . . .	694

## Table des matières du volume Applications et les Composants

# APPLICATIONS

## PRODUCTION, TRANSPORT ET DISTRIBUTION DE L'ÉNERGIE ÉLECTRIQUE

Les applications de la commande automatique dans les systèmes de production, de transport et de distribution de l'énergie électrique—  
Une synthèse . . . . . *B. Favez*

### Production, transport et distribution de l'énergie électrique

Mise au point d'une méthode pour la commande cybernétique de systèmes énergétiques électriques intégrés  
*V. A. Venikov et L. V. Tsukernik*

Quelques résultats récents dans la commande de systèmes énergétiques au moyen de calculateurs  
*T. Vámos, S. Benedikt et M. Uzsoki*

Commande optimale du fonctionnement de systèmes thermohydrauliques . . . . . *L. K. Kirchmayer et R. J. Ringlee*

Systèmes automatiques avec éléments à apprentissage . . . . . *G. K. Krug et A. V. Netushil*

Principes et réalisations des automatismes liés à la manutention de combustible d'un réacteur nucléaire . . . . . *P. Turpin et J. Thilliez*

Une étude de la dynamique de formation des bulles de vapeur dans les réacteurs nucléaires à eau bouillante  
*P. K. M'Pherson et M. Muscettola*

### Le réglage des systèmes hydrauliques

Le réglage optimal de systèmes de production hydro-électriques . . . . . *H. Ruge*

Exposé d'une méthode d'élaboration de graphiques exprimant les conditions de stabilité du réglage d'un groupe hydro-électrique  
*A. Tschumy*

Étude au moyen de calculateurs numériques de systèmes de production d'énergie à machines multiples . . . . . *H. Glavitsch*

### Commande des machines de production d'énergie électrique

Commande optimale de turbines hydrauliques compte tenu des non-linéarités de la vitesse du servo-moteur . . . . . *Th. Stein*

Equations de commande d'une centrale hydro-électrique avec des valeurs de référence fixes . . . . . *L. Borel*

## INDUSTRIE SIDERURGIQUE

Applications de l'automatisation et des techniques au laminage et au traitement des métaux — Une synthèse . . . . . *W. E. Miller*

Résultats obtenus par l'automatisation et la sidérurgie — Une synthèse . . . . . *A. Ya. Lerner*

Commande de la préparation de mélanges à agglomérer . . . . . *G. DeGregorio, G. Litigio et G. Sironi*

Planification dynamique d'une aciérie avec fours à sole . . . . . *M. Korobko et Yu. Samoilenko*

L'automatisation du forgeage de grosses pièces . . . . . *J. G. Wistreich et A. Tomlinson*

L'automatisation d'une aciérie avec référence particulière à l'emploi de calculateurs numériques pour la préparation de la production  
et la transmission des informations . . . . . *S. E. Hersom et R. G. Massey*

Commande d'un laminoir de finissage de bandes d'acier à chaud au moyen de calculateurs en ligne . . . . . *R. G. Beadle*

Commande optimale de processus continus . . . . . *A. Ya. Lerner*

Un système optimal numérique de commande programmée et son application au mécanisme de serrage d'un laminoir de blooming  
*S. M. Domanitsky, V. V. Imedadze et Sh. A. Tsintsadze*

Système de commande du processus de recuit continu au moyen d'un calculateur . . . . . *J. T. Bradford, Jr.*

## LES INDUSTRIES DE LA CHIMIE ET DU PÉTROLE

Une synthèse de l'application aux industries de la chimie et du pétrole . . . . .	<i>H. W. Slotboom, J. de Jong, J. E. Rijnsdorp</i>
Caractéristiques dynamiques d'un colonne de distillation binaire . . . . .	<i>K. Izawa et T. Morinaga</i>
Modèles approximatifs de la reponse dynamique de grandes colonnes de distillation . . . . .	<i>J. S. Moczek, R. E. Otto et T. J. Williams</i>
Une étude du comportement dynamique d'un système de récupération d'énergie de cracking catalytique au moyen d'un calculateur analogique . . . . .	<i>C. A. J. M. van der Heyden et A. G. van Nes</i>
Analyse statistique d'un nouveau système de réglage de débits de fluides . . . . .	<i>R. C. Booton, Jr. et W. E. Sollecito</i>
Effets du mélange de fluides et de ses expressions sur la dynamique du processus de transfert des masses . . . . .	<i>T. Takamatsu et E. Nakanishi</i>
Etude expérimentale du comportement dynamique d'un échangeur de chaleur et d'un processus de mélange . . . . .	<i>L. Delvaux</i>
Etude de la production industrielle du polyéthylène sous des hautes pressions et du réglage automatique de ce processus . . . . .	<i>B. V. Volter</i>
Etude des caractéristiques statiques et dynamiques de processus de distillation fractionnée . . . . .	<i>I. V. Anisimov</i>
Aptitude au réglage et capacité admissible du compresseur d'un système de récupération d'un gaz inflammable . . . . .	<i>F. J. Kylstra</i>
Analyse et étude d'un système adaptatif à perturbation de paramètres pour son application au réglage de processus . . . . .	<i>T. Isobe et T. Totani</i>
Propriétés dynamiques de stations de rectification avec des colonnes à plateaux . . . . .	<i>J. Závorka</i>
Une réalisation originale dans une raffinerie de pétrole: le chargement automatique des wagons citernes . . . . .	<i>F. X. Montjean</i>

## AUTOMATISATION DES PROCESSUS INDUSTRIELS

Le traitement du problème d'optimisation par A 110 . . . . .	<i>E. Honoré</i>
Automatisation d'une usine de ciment Portland au moyen d'un calculateur numérique de commande . . . . .	<i>R. A. Phillips</i>
Sur la stabilité et l'étude de systèmes adaptatifs à oscillations exploratrices . . . . .	<i>R. K. Smyth et N. Nahi</i>

## Réglages des systèmes à vapeur

Predétermination des résultats du réglage des résurchauffeurs de générateur de vapeur . . . . .	<i>W. Kindermann</i>
Une optimisation du rendement de chaudières . . . . .	<i>S. Fujii et N. Kanda</i>
Un simulateur de turbines à vapeur avec résurchauffe ou extraction réglée automatiquement . . . . .	<i>H.-J. Ehling</i>
Un système de réglage électro-hydraulique de turbines à résurchauffe . . . . .	<i>M. A. Eggenberger et P. H. Troutman</i>

## Industries des moteurs

Une application industrielle d'un calculateur intérieur à un circuit de commande: l'équipement électronique de commande d'une machine à équilibrer les vilebrequins . . . . .	<i>J. Csech</i>
Analyse d'un système de réglage automatique de la vitesse d'un moteur . . . . .	<i>W. H. Holl</i>

## SYSTÈMES COMBINES HOMME-MACHINE

Modèles continus et échantillonnés de l'opérateur humain placé dans une boucle de commande . . . . .	<i>P. Naslin et J.-C. Raoult</i>
Modèles discrets de l'opérateur humain dans un système de commande . . . . .	<i>G. E. Bekey</i>
Analyse dynamique et simulation des fonctions de commande d'une direction d'entreprise . . . . .	<i>R. B. Wilcox</i>
Esquisse d'une théorie de commande automatique de prothèses . . . . .	<i>R. Tomovic</i>
Un modèle échantillonné des mouvements de poursuite de l'oeil . . . . .	<i>L. R. Young</i>

## TECHNIQUES D'APPLICATION

## Méthodes

Méthode auto-adaptative permettant de tenir compte de grandes variations du gain du système réglé dans les systèmes de commande . . . . .	<i>R. J. Kochenburger</i>
Dynamique des lignes de transmission hydrauliques . . . . .	<i>R. Oldenburger et R. E. Goodson</i>
L'optimisation des systèmes de commande par calculateur utilisant une connaissance partielle de l'état de la grandeur de sortie . . . . .	<i>B. G. Anderson</i>

## AUTOMATIQUE DES SYSTÈMES AERO-SPATIAUX

## Systèmes aéro-spatiaux

Etude d'un système de réglage de la pression pour une soufflante de 1,50 m sur 1,50 m . . . . .	<i>J. A. Tanner et W. Dietiker</i>
Simulation de l'ambiance de l'engin spatial pour une installation d'essais de moteurs de fusées . . . . .	<i>G. J. Fiedler et J. J. Landy</i>
Système de pilotage automatique longitudinal d'atterrissage d'avion sans visibilité . . . . .	<i>F. J. Ellert et C. W. Merriam III</i>
Commande automatique d'une grande antenne mobile de télécommunication avec des satellites . . . . .	<i>F. J. D. Taylor</i>
Etude d'un système de commande automatique d'un radio-télescope de 63 m. . . . .	<i>R. G. Wheeler</i>
Modèle dynamique pour la commande d'altitude à visée fine de l'observatoire astronomique placé sur une orbite . . . . .	<i>R. E. Roberson</i>
Relations fondamentales de réponse pour la commande d'altitude de satellites utilisant des gyroscopes . . . . .	<i>R. H. Cannon, Jr.</i>

## RAPPORT

<i>W. Miller</i> (Président du Comité Technique des Applications) . . . . .	
---	--

## LES COMPOSANTS

Nouveaux principes d'étude et dispositifs de commande — Une synthèse réalisée . . . . .	<i>J. L. Shearer et 16 autres auteurs</i>
---	---

## DISPOSITIFS MECANIKES, HYDRAULIQUE DE COUPLE

Un amplificateur hydraulique de couple . . . . .	<i>Y. Oshima et K. Araki</i>
Un instrument gyroscopique à entraînement rotatif et à signal de sortie vibratoire . . . . .	<i>G. C. Newton</i>
Quelques problèmes de la dynamique d'un servo-mécanisme hydraulique à commande par laminage et charge inertielle . . . . .	<i>V. Khokhlov</i>
Réalisation de machines séquentielles avec des moyens pneumatiques . . . . .	<i>A. A. Tal</i>

## DISPOSITIFS ÉLECTRO-MÉCANIQUES ET AMPLIFICATEURS MAGNÉTIQUES

Dispositif fréquentiel fonctionnel à deux positions pour le réglage automatique . . . . .	<i>I. A. Maslaroff</i>
Les problèmes, le fonctionnement et le calcul d'un nouveau composant destiné à être appliqué dans certains circuits de commande . . . . .	<i>O. Benedikt</i>
Le commande d'un mécanisme oscillant électro-magnétique linéaire . . . . .	<i>J. C. West et B. V. Jayawant</i>
Moteurs pas-à-pas à rotor actif . . . . .	<i>Yu. K. Vasilev, Yu A. Prokofiev et G. Ya. Vainberger</i>

## COMPOSANTS ÉLECTRONIQUES

Quelques nouveaux circuits de commande utilisant des triodes semiconductrices à quatre couches du type $p-n-p-n$ . . . . .	<i>J. Š. Haškovec et A. Klímek</i>
Redresseurs commandés au silicium à commutation . . . . .	<i>H. F. Storm</i>
Mémoires céramiques dans les ambiances extrêmes . . . . .	<i>A. B. Kaufman</i>

## DISPOSITIFS NUMÉRIQUES

L'utilisation des analyseurs différentiels numériques dans les boucles de réglage . . . . .	<i>H. Rechberger</i>
Avantages et possibilités du réglage numérique des vitesses . . . . .	<i>W. Fritzsche</i>
Les régulateurs numériques . . . . .	<i>T. M. Aleksandridi, K. S. N. Diligensky et H. K. Krug</i>

## INSTRUMENTATION DE PROCESSUS

## Composants divers

Un analyseur statistique universel . . . . .	<i>J. Krýže</i>
Le comportement des régulateurs adaptatifs . . . . .	<i>J. L. Douce</i>
Les principes d'étude et le circuit d'un corrélateur à canaux multiples — un calculateur analogique spécialisé pour le traitement statistique des séries temporelles aléatoires dans les systèmes de commande industrielle . . . . .	<i>A. S. Uskov et Yu. M. Orlov</i>

## FIABILITÉ DES COMPOSANTS

La fiabilité des composants — Une synthèse . . . . . *G. S. Glinski, B. S. Sotskov et H. Weissmann*

**La fiabilité**

Nouvelles servo-valves pour la commande électro-hydraulique surabondante . . . . . *K. D. Garnjost et W. J. Theyer*

La fiabilité des composants électroniques . . . . . *G. W. A. Dummer*

Problèmes de la fiabilité d'éléments électro-mécaniques . . . . . *B. S. Sotskov, I. E. Dekabrun et L. S. Krivorotova*

Une étude de la fiabilité des servomécanismes dans les systèmes de commande des réacteurs et des centrales nucléaires . . . . .  
*L. A. J. Lawrence et R. J. Scotcher*

## RAPPORTS

*Gy. Boromisza* (Président du Comité Technique des Composants) . . . . .

*Y. Oshima* (Vice-Président du Comité Technique des Composants) . . . . .

# INHALT

MITTEILUNG DER HERAUSGEBER . . . . .	V
VERZEICHNIS DER VERFASSER . . . . .	XXXI

## ANSPRACHEN

BEGRÜSSUNG . . . . .	PROFESSOR ED. GERECKE (Präsident der International Federation of Automatic Control)	XXXV
WIE ES ZUR GRÜNDUNG DES IFAC KAM . . . . .	H. CHESTNUT (1. Präsident des IFAC, 1957-1959)	XXXVI
DER 1. IFAC-KONGRESS . . . . .	PROFESSOR A. M. LETOV (2. Präsident des IFAC, 1959-1961)	
DIE ENTWICKLUNG DER NACHRICHTENÜBERTRAGUNG UND IHR EINFLUSS AUF DIE REGELUNG . . . . .	J. L. AUERBACH (Präsident der International Federation of Information Processing Societies)	XXXVIII

## THEORIE

### THEORIE NICHTLINEARER SYSTEME

Statistische Verfahren zur Behandlung von Regelvorgängen — Übersichtvortrag . . . . .	V. S. Pugachev	1
---	----------------	---

#### Technik der Beschreibungsfunktion

Eine neue Methode zur Ableitung der Beschreibungsfunktion bestimmter nichtlinearer Übertragungssysteme . . . . .	R. Lauber	14
Anwendung des Verfahrens der Beschreibungsfunktion auf die Untersuchung von Schwingungen in Systemen mit veränderlichen Parametern . . . . .	A. Leonhard	21
Über das Problem der inversen Beschreibungsfunktion . . . . .	J. E. Gibson und E. S. di Tada	29
Die relative Stabilität von Schwingungen in nichtlinearen Regelsystemen . . . . .	Z. Bonenn	35

#### Prädiktionssysteme

Prädiktionsregelung eines Zweipunktsystems mit zwei Stellgrößen . . . . .	A. J. Adey, J. F. Coales und J. A. Styles	41
Ein Prädiktionsverfahren für nichtstationäre Prozesse und seine Anwendung auf das Problem der Vorhersage der Lastverteilung im Verbundnetz . . . . .	E. D. Farmer	47
Über den Entwurf von Prädiktions-Regelsystemen . . . . .	S. Horing	55
Ein Verfahren zur zeitoptimalen Regelung mit Vorhersage . . . . .	F. B. Gul'ko und B. Y. Kogan	63

#### Nichtlineare Systeme mit zufallsbedingten Parametern

Optimierung nichtlinearer stochastischer Regelvorgänge . . . . .	R. Kulikowski	69
Nichtlineares Programmieren bei der Untersuchung optimaler Regelsysteme . . . . .	N. I. Andreew	76
Numerische Analyse nichtlinearer Regelsysteme mit Hilfe der Gleichung von Fokker, Planck und Kolmogorow . . . . .	K. J. Merklinger	81
Eine Darstellung zeitveränderlicher nichtlinearer Systeme mit Hilfe einer Volterra'schen Reihe . . . . .	R. H. Flake	91

## Nichtlineare stochastische Systeme

Die Anwendung quasi-linearer Methoden auf nichtlineare Regelsysteme mit zufallsbedingten Eingangsgrößen . . . . .	<i>H. W. Smith</i>	100
Ein Digital-Rechenverfahren zur Untersuchung nichtlinearer Systeme mit Anwendung auf zufallsbedingte Prozesse . . . . .	<i>T. Prasad und V. P. Sinha</i>	108
Zeitoptimale Systeme mit stochastischen Störgrößen . . . . .	<i>V. N. Novoseltsev</i>	114

## Stückweise kontinuierliche Systeme

Mehrfach-Regelkreise mit gemischt analog-digitaler Signalverarbeitung und beschränkten Störgrößeneinflüssen . . . . .	<i>V. Strejc</i>	122
Optimalregelung eines nichtlinearen Prozesses mit Vorschriften über diskrete Zeitabschnitte . . . . .	<i>B. Friedland</i>	128
Theorie der schnelligkeitsoptimalen Regler . . . . .	<i>C. A. Desoer, E. Polak und J. Wing</i>	135

## Abtastsysteme

Über die Wurzeln eines reellen Polynoms innerhalb des Einheitskreises und ein Stabilitätskriterium für in diskreten Zeitabschnitten lineare Systeme . . . . .	<i>E. I. Jury</i>	142
Analytische Verfahren zur Berechnung nichtlinearer Abtastsysteme . . . . .	<i>B. Kondo und S. Iwai</i>	154
Ein stetigwirkendes Kompensationsfilter für lineare Abtastsysteme . . . . .	<i>B. M. Brown</i>	165
Grundlagen der Theorie nichtlinearer Regelsysteme mit Abtastung . . . . .	<i>Ya. Z. Zypkin</i>	172
Synthese optimaler Abtastsysteme . . . . .	<i>L. N. Volgin</i>	181
Regelkreise mit Integral-Pulsfrequenz-Modulator . . . . .	<i>C. C. Li und R. W. Jones</i>	186
Die Kombination endlicher Beruhigungszeit und minimalen quadratischen Fehlers in digitalen Regelsystemen . . . . .	<i>V. Peterka</i>	196

## Relais-Systeme

Subharmonische Schwingungen in einem Zweipunkt-Regelsystem . . . . .	<i>J. C. Gille, S. Wegrzyn und J. C. Paquet</i>	204
Die Synthese von Regelkreisen mit linearem Verhalten bei kleinen, mit Zweipunkt-Verhalten bei großen Signalen . . . . .	<i>E. V. Persson</i>	210
Systeme mit zwei Eingängen und Beschränkung in Gestalt einer Sättigung . . . . .	<i>R. S. Gaylord</i>	219

## Theorie der Logik-Schaltungen

Warnung und Vorhersage von Versagern in Steuerungssystemen mit Struktur-Redundanz . . . . .	<i>M. A. Gavrilo</i>	228
Axiomatik der Theorie der Vereinfachung von Schaltkreisen . . . . .	<i>Gr. C. Moisi</i>	234
Adaptierende Regelung eines Systems mit einer endlichen Anzahl von Zuständen . . . . .	<i>S. Paszhkovskij</i>	241

# OPTIMALE REGLER

Synthese optimaler Regler — Übersichtsvortrag . . . . .	<i>A. M. Letov</i>	246
---	--------------------	-----

## Optimale Systeme

Näherungsverfahren für Optimalwert- und Adaptiv-Regelung . . . . .	<i>J. H. Westcott, J. J. Florentin und J. D. Pearson</i>	263
Eine optimale Leitbahnnäherung für fast kreisförmige Begegnungsbahnen von Flugkörpern . . . . .	<i>H. J. Kelly und J. C. Dunn</i>	274
Zur Synthese optimaler Regelsysteme . . . . .	<i>L. W. Neustadt und B. H. Paiewonsky</i>	283
Eine Anwendung der Optimalwertregelung auf die mittlere Flugphase von Flugkörpern . . . . .	<i>J. S. Meditch und L. W. Neustadt</i>	292
Über notwendige und hinreichende Bedingungen für geschwindigkeitsoptimale Regelung nichtlinearer Systeme mit Verzögerung 2. Ordnung . . . . .	<i>E. B. Lee und L. Markus</i>	300
Über die zeitoptimale Regelung von Systemen mit mehreren Eingangsgrößen und unterschiedlichen Beschränkungen . . . . .	<i>P. E. Sarachik und G. M. Kranc</i>	306
Die Näherungsberechnung einer Klasse von Regelsystemen mit selbsteinstellendem Netzwerk . . . . .	<i>Y. I. Alimov</i>	315

## Optimalsysteme mit verteilten Parametern

Einige Betrachtungen zur optimalen Regelung eines vermaschten Systems . . . . .	<i>R. Marcacci</i>	324
Optimale Abläufe in Systemen mit Signalverzögerungen . . . . .	<i>N. N. Krasovsky</i>	327
Die Optimalwertregelung von Systemen mit verteilten Parametern . . . . .	<i>A. G. Butkovsky</i>	333



**Synthese optimaler Systeme**

Die Lösung eines Problems der geschwindigkeitsoptimalen Regelung mit Hilfe des Maximumprinzips von Pontryagin	<i>Y. Sakawa und Ch. Hayashi</i>	339
Analyse und Synthese zeitoptimaler Regelsysteme	<i>Sun Jian und Hang King-ching</i>	347
Entwicklung von Verfahren der dynamischen Programmierung und ihre Anwendung auf den Entwurf optimaler Systeme	<i>R. L. Stratonovich</i>	352
Ein abgewandeltes Maximumprinzip zur optimalen Regelung eines Systems mit begrenzten Phasenraum-Koordinaten	<i>S. S. L. Chang</i>	358
Optimale und fast-optimale Regelung von Systemen 3. und 4. Ordnung	<i>I. Flügge-Lotz und H. A. Titus, Jr.</i>	363

**Optimales Programmieren**

Programmsteuerung und Theorie von Optimalsystemen	<i>Ye. A. Barbashin</i>	371
Die Verwirklichung von Optimalprogrammen für Regelsysteme	<i>G. S. Pospelov</i>	377
Über die Grenzen von Quantisierungsfehlern bei Berechnungen zur dynamischen Programmierung	<i>J. J. G. Guignabodet</i>	383

**THEORIE DER SELBSTEINSTELLENDEN SYSTEME**

Adaptierende und selbstoptimierende Systeme — Übersichtsvortrag	<i>J. Truxal</i>	386
Lernende Systeme — Übersichtsvortrag	<i>G. Pask</i>	393

**Invarianzprobleme**

Über die Invarianzbedingungen bei gewöhnlichen und adaptierenden Abtastsystemen	<i>V. M. Kunzewitsch und Yu. V. Krementulo</i>	412
Optimierung aufgrund von Invarianzbedingungen in Regelsystemen mit fester und veränderlicher Struktur	<i>B. N. Petrow, G. M. Ulanov und S. V. Emelyanov</i>	421
Synthese von Systemen, bei denen das entsprechende selbsteinstellende Ersatzmodell feste Kennwerte besitzt	<i>M. V. Mejerow</i>	430

**Adaptierende Systeme**

Ein Vergleich der Auswertezeit in selbsteinstellenden Regelkreisen	<i>F. Mesch</i>	439
Selbsteinstellende Regelsysteme ohne Eingangs-Testsignale	<i>E. P. Popov, G. M. Loskutov und R. M. Yusupov</i>	446
Ein Beitrag zur Anwendung selbsteinstellender Systeme auf die automatisierte Synthese von Lernsystemen	<i>V. K. Chichinadze</i>	453
Das Auffinden von Extrema bei Funktionen, die in Regelkreisen auftreten	<i>A. A. Woronow und M. B. Ignatjev</i>	459
Versuch der Anwendung eines dynamischen Leitmodells auf die Theorie adaptierender Systeme	<i>A. Straszak</i>	465
Verallgemeinerte Analyse selbsteinstellender Systeme, die auf einem Suchverfahren mit sinusförmigen Testsignalen beruhen	<i>V. W. Eveleigh</i>	472
Über die Durchführung eines Programmes der Adaptivregelung in einem System mit Digitalrechner	<i>P. F. Klubnikin</i>	481

**Lernende Systeme**

Ein adaptierendes Regelsystem mit „Formempfindung“	<i>W. K. Taylor</i>	488
STELLA: Ein Konzept für einen lernenden Automaten	<i>J. H. Andreae</i>	497
Lernfähige Systeme für technische Anwendungen — unter Berücksichtigung von Versuchen über die Möglichkeit, die Geräte ein Erkennen von Mustern zu lehren	<i>M. A. Aizerman</i>	503

**Suchverfahren**

Über die Theorie selbsteinstellender Systeme mit Gradientenbestimmung, nach dem Verfahren des Hilfsoperators	<i>I. E. Kasakow und L. G. Evlanow</i>	510
Über die Anwendung der Theorie kontinuierlicher Systeme auf die Extremwert-Regelung in industriellen Anlagen	<i>A. A. Krasowski</i>	519
Prinzip und Anwendung eines Optimalwert-Rechners	<i>R. Perret und R. Rouxel</i>	527
Probleme der „dualistischen“ Theorie der Optimalwert-Regelung	<i>A. A. Feldbaum</i>	541
Die Anwendung stochastischer Testsignale bei der Optimierung von Prozessen	<i>P. M. E. M. van der Grinten</i>	551

## METHODEN ZUR ABSCHÄTZUNG DER STABILITÄT

## Das Stabilitätsproblem nach Ljapunow

Über eine neuartige Erweiterung des Stabilitätsbegriffes . . . . .	<i>J. P. La Salle und R. J. Rath</i>	556
Stabilitätsuntersuchung, mittels der 2. Methode von Ljapunow, an Systemen mit beschränkten Nichtlinearitäten . . . . .	<i>H. Nour-Eldin</i>	561
Die Anwendung der Technik der linearen Bereichsgrenzen zur Übertragung der direkten Ljapunow'schen Methode auf eine Gruppe von nichtlinearen und zeitveränderlichen Systemen . . . . .	<i>R. A. Nesbit</i>	568
Über die Abschätzung der Abklingzeit . . . . .	<i>L. Hwang</i>	576
Neue Verfahren zur Ermittlung Ljapunow'scher Funktionen für zeit-invariante Regelsysteme . . . . .	<i>G. P. Szegö</i>	584
Ein Verfahren zur Stabilitätsprüfung . . . . .	<i>H. H. Rosenbrock</i>	590

## DYNAMIK DER SYSTEME UND VERWANDTE PROBLEME

Prozeßdynamische Studien und ihre Anwendung auf Entwurf und Regelung industrieller Anlagen — Übersichtsvortrag

*T. J. Williams* 595

## Dynamische Probleme

Ermittlung des dynamischen Verhaltens von Regelkreisen mit Hilfe einstellbarer Modelle . . . . .	<i>E. Blandhol und J. G. Balchen</i>	602
Die systematischen und zufallsbedingten Fehler bei der experimentellen Bestimmung von Übertragungsfunktionen . . . . .	<i>J. Loeb</i>	615
Die Anwendung des Relais-Korrelators und des Polaritäts-Korrelators auf Messungen an Regelkreisen . . . . .	<i>B. P. Th. Veltman und A. van den Bos</i>	620
Bemerkungen zu einer Zufallsfunktion mit physikalischer Bedeutung . . . . .	<i>M. J. Pélégryn</i>	628
Eine Unschärfe-Beziehung für lineare mathematische Modelle . . . . .	<i>B. Qvarnström</i>	634

## Allgemeine Probleme

Axiomatische Begründung der Theorie geregelter Systeme . . . . .	<i>E. Roxin</i>	640
Das Umkehrproblem zur Bestimmung des quadratischen Integral-Kriteriums . . . . .	<i>W. Jaromínek</i>	645
Über Systeme mit Regelung der Konfiguration . . . . .	<i>J. Beneš</i>	656
Annäherung industrieller Regelsysteme an optimale nichtlineare Regelungen mit begrenzter Stellgeschwindigkeit mittels konventioneller Regler . . . . .	<i>E. Pavlík</i>	664
Neues Verfahren zur Optimierung bei stochastischen Eingangsgrößen mittels der Transformation $V = (Z - 1)/(Z + 1)$ . . . . .	<i>P. M. Lefèvre</i>	671
Die Regelung von Anlagen über zwei Bereiche der Störampplituden . . . . .	<i>C. N. Kerr und G. D. S. MacLellan</i>	682

## SCHLUSSBEMERKUNGEN

<i>B. N. Petrow</i> (Vorsitzender des Technischen Ausschusses im IFAC für Theorie) . . . . .	693
<i>J. H. Westcott</i> (Stellvertretender Vorsitzender des Ausschusses im IFAC für Theorie) . . . . .	694

## Inhalt des Bandes Anwendungen und Geräte

# ANWENDUNGEN

## AUF DEM GEBIET DER ELEKTRISCHEN ENERGIEVERSORGUNG

Die Regelung in elektrischen Netzen — Übersichtsvortrag . . . . . *B. Favez*

### Elektrizitätswerke und Lastverteilung

Entwicklung eines Verfahrens zur Regelung großer Verbundnetze nach kybernetischen Gesichtspunkten  
*V. A. Venikow und L. V. Tsukernik*  
Einige neuere Ergebnisse in der Steuerung der Lastverteilung durch Rechner . . . . . *T. Vámos, S. Benedikt und M. Uzsoki*  
Optimalregelung eines Verbundnetzes von Wärme- und Wasserkraftwerken . . . . . *L. K. Kirchmayer und R. J. Ringlee*

### Energieerzeugung

Über Regelsysteme mit lernfähigen Elementen . . . . . *G. K. Krug und A. V. Netushil*  
Prinzip und Konstruktion von Mechanismen zur Steuerung der Brennelemente in einem Kernreaktor *P. Turpin und J. Thilliez*  
Eine Untersuchung der Dynamik der Blasenverteilung in Siedewasserreaktoren . . . . . *P. K. M'Pherson und M. Muscettola*

### Regelung im Wasserkraftwerk

Über die optimale Regelung von Wasserkraftwerken . . . . . *H. Ruge*  
Vorschlag für ein Verfahren zur Ermittlung der Stabilitätsbedingungen der Regelung eines Wasserkraftwerks . . . . . *A. Tschumy*  
Die Untersuchung des Verhaltens von Generatorgruppen mittels eines Digitalrechners . . . . . *H. Glavitsch*

### Regelung der Kraftmaschinen im Elektrizitätswerk

Optimale Auslegung der Regler für Wasserturbinen unter Berücksichtigung der Nichtlinearität im Stellmotor . . . . . *Th. Stein*  
Die Regelgleichungen eines Wasserkraftwerkes bei festen Bezugswerten . . . . . *L. Borel*

## STAHLINDUSTRIE

Anwendung der Automatisierung auf den Walzwerksbetrieb — Übersichtsvortrag . . . . . *W. E. Miller*  
Fortschritte in der Automatisierung der Stahlindustrie — Übersichtsvortrag . . . . . *A. Ya. Lerner*

### Stahlindustrie

Regelung beim Aufbereiten einer Sintermischung . . . . . *G. DeGregorio, G. Litigio und G. Sironi*  
Dynamisches Programmieren im Siemens-Martin-Ofen . . . . . *M. Korobko und Yu. Samoilenko*  
Automatisierung von Schmiedepressen . . . . . *J. G. Wistreich und A. Tomlinson*  
Die Automatisierung in einem Hüttenwerk mit dem Einsatz eines Digitalrechners zur Steuerung des Produktionsablaufs und der Datenübertragung . . . . . *S. E. Hersom und R. G. Massey*  
Über den Einsatz eines „On-line“-Prozeßrechners in einer Warmband-Fertigungsstraße. . . . . *R. G. Beadle*  
Optimalwertregelung von Bandprozessen . . . . . *A. Ya. Lerner*  
Eine digitale Programmsteuerung und ihre Anwendung auf die Regelung der Walzenanstellung in einer Block-Brammenstraße  
*S. M. Domanitsky, V. V. Imedadze und Sh. A. Tsintsadze*  
Die Steuerung einer kontinuierlichen Glühstraße durch einen Prozeßrechner . . . . . *J. T. Bradford, Jr.*

## CHEMISCHE UND ERDÖL-INDUSTRIE

Eine Übersicht über die Anwendung der Regelung in der chemischen Großindustrie und Erdölaufbereitung

*H. W. Slotboom, J. de Jong und J. E. Rijnsdorp*

**Chemische und Erdöl-Industrie**

- Dynamische Kenngrößen der Zweistoff-Trennkolonne . . . . . *K. Izawa und T. Morinaga*  
 Näherungsmodelle für das dynamische Verhalten großer Trennkolonnen . . . . . *J. S. Moczek, R. E. Otto und T. J. Williams*  
 Über das dynamische Verhalten des Systems zur Energierückgewinnung einer Krackanlage mit Hilfe eines Analogrechners  
*C. A. J. M. van der Heyden und A. G. van Nes*  
 Entwicklung eines neuartigen Verfahrens zur Messung des Durchflusses von Flüssigkeiten . . . *R. C. Booton, Jr. und W. E. Sollecito*  
 Über Effekte, die bei der Mischung von Flüssigkeiten auftreten, und ihre Auswirkung auf das dynamische Verhalten von Anlagen mit Stofftransport . . . . . *T. Takamatsu und E. Nakanishi*  
 Experimentelle Untersuchungen zum dynamischen Verhalten eines Wärmetauschers und eines Misch-Prozesses . . . *L. Delvaux*  
 Untersuchung einer Anlage zur Polymerisation von Äthylen unter hohem Druck und die zugeordneten Regelprobleme *B. V. Volter*  
 Eine Berechnung der dynamischen und statischen Kenngrößen von Anlagen zur fraktionierten Destillation . . . *I. V. Anisimow*  
 Über die Regelfähigkeit und zulässige Kompressorleistung einer Anlage zur Abgas-Wiedergewinnung . . . . . *F. J. Kylstra*  
 Analyse und Entwurf eines für die Verfahrensregelung geeigneten, selbsteinstellenden Systems mit Parametersteuerung  
*T. Isobe und T. Totani*  
 Die dynamischen Eigenschaften von Rektifizierkolonnen . . . . . *J. Závorka*  
 Über eine Verbesserung im Betrieb einer Erdöl-Raffinerie: Die automatische Beladung von Kesselwagen . . . . *F. X. Montjean*

## AUTOMATISIERUNG IN DER VERFAHRENSTECHNIK

**Prozeßrechner**

- Die Lösung eines Optimierungsproblems mit Hilfe des Digitalrechners A. 110 . . . . . *E. Honoré und Ste. Analae*  
 Die Automatisierung einer Portland-Zementfabrik mittels eines Prozeßrechners . . . . . *R. A. Phillips*  
 Über die Stabilität und Auslegung von selbsteinstellenden Systemen mit selbsterregter Suchschwingung . . . *R. K. Smyth und N. Nahi*

**Regelung im Dampfkraftwerk**

- Vorausberechnung des Regelverhaltens von Zwischenüberhitzern im Dampfkessel . . . . . *W. Kindermann*  
 Eine Optimalwertregelung des Kesselwirkungsgrades . . . . . *S. Fujii und N. Kanda*  
 Ein Simulator für Dampfturbinen mit regelbarer Entnahme . . . . . *H.-J. Ehling*  
 Ein elektro-hydraulisches Regelsystem für Turbinen mit Zwischenüberhitzern . . . . . *M. A. Eggenberger und P. H. Troutman*

**Kraftfahrzeug-Industrie**

- Industrielle Anwendung eines Prozeßrechners auf eine Fertigungseinrichtung: Elektronisches Steuersystem für eine Maschine zum Auswuchten von Kurbelwellen . . . . . *J. Csech*  
 Entwurf eines Gerätes zur Regelung der Geschwindigkeit eines Kraftwagens . . . . . *W. H. Holl*

## SYSTEME MIT REGELUNG UNTER MITWIRKUNG DES MENSCHEN

**Der Mensch als Regelkreisglied**

- Kontinuierliche und Abtastmodelle für das Verhalten des Menschen im Gefüge eines Regelsystems . . . *P. Naslin und J.-C. Raoult*  
 Zeitdiskrete Modelle für den Menschen als Regelkreisglied . . . . . *G. E. Bekey*  
 Analyse der Dynamik und Nachbildung der Steuerungsfunktionen in der Betriebsführung . . . . . *R. B. Wilcox*

**Physiologische und medizinische Probleme**

- Grundzüge einer Theorie steuerbarer Prothesen . . . . . *R. Tomović*  
 Ein Abtastmodell zur Beschreibung der Zielbewegung des Auges . . . . . *L. R. Young*

## WEITERE REGELVERFAHREN UND GERÄTE

### Verfahren

- Ein adaptierendes Verfahren zum Ausgleich großer Änderungen des Übertragungsfaktors einer Regelstrecke . . . *R. J. Kochenburger*  
 Das dynamische Verhalten von Hydraulik-Leitungen . . . . . *R. Oldenburger und R. E. Goodman*  
 Die Optimierung eines durch einen Prozeßrechner gesteuerten Systems bei lückenhafter Information über die inneren funktionellen Zusammenhänge . . . . . *B. G. Anderson*

## REGELUNG IN DER LUFT- UND RAUMFAHRT

### Prüf- und Meßeinrichtungen

- Entwurf eines Druckregelsystems für einen Windkanal von  $1,5 \cdot 1,5 \text{ m}^2$  Strömungsquerschnitt . . . *J. A. Tanner und W. Dietiker*  
 Ein Prüfstand für Raketenmotoren mit Nachbildung der realen Umweltbedingungen . . . . . *G. J. Fiedler und J. J. Landy*  
 Ein Leitsystem für die Längsbewegung eines Flugzeugs in der Landephase . . . . . *F. J. Ellert und C. W. Merriam III*

### Hilfseinrichtungen zur Steuerung von Satelliten

- Lageregelung einer großen Richtantenne für Fernmelde-Satelliten . . . . . *F. J. D. Taylor*  
 Untersuchung der Nachführeinrichtung eines 64-m-Radioteleskops . . . . . *R. G. Wheeler*  
 Ein dynamisches Modell der Feineinstellung eines astronomischen Beobachtungssatelliten . . . . . *R. E. Roberson*  
 Probleme der Lageregelung von Raumfahrzeugen mit Hilfe von Kreiselssystemen . . . . . *R. H. Cannon, Jr.*

## SCHLUSSBEMERKUNGEN

- W. E. Miller* (Vorsitzender des Technischen Ausschusses im IFAC über Anwendungsfragen) . . . . .

## GERÄTE

Neue Konstruktionsgedanken und Regelgeräte — Übersichtsreferat . . . . . *J. L. Shearer und 16 weitere Autoren*

## MECHANISCHE, HYDRAULISCHE UND PNEUMATISCHE GERÄTE

Ein hydraulischer Drehmomentverstärker . . . . . *Y. Oshima und K. Araki*  
 Ein Kreiselgerät mit Spezialrotor und periodischem Ausgangssignal . . . . . *G. C. Newton*  
 Über dynamische Probleme einer hydraulischen Steuereinrichtung bei Belastung mit schweren Massen . . . . . *V. Khokhlov*  
 Der Aufbau von Sequenzsteuerungen mit Hilfe pneumatischer Schaltelemente . . . . . *A. A. Tal*

## ELEKTROMECHANISCHE GERÄTE UND MAGNETVERSTÄRKER

Zweipunktregelung mit gesteuerter Frequenz der Arbeitsbewegung . . . . . *I. A. Maslaroff*  
 Probleme, Wirkungsweise und Berechnung eines neuen Maschinenverstärkers und seine Anwendung auf bestimmte Regelaufgaben  
*O. Benedikt*  
 Die Regelung der Bewegung eines translatorischen elektromagnetischen Schwingantriebs . . . . . *J. C. West und B. V. Jayawant*  
 Schrittmotoren mit aktivem Läufer . . . . . *Yu. K. Vasilev, Yu. A. Prokofiev und G. Ya. Wainberger*

## ELEKTRONISCHE BAUELEMENTE

Einige neuartige Steuerschaltungen mit Vierschicht-*p-n-p-n*-Halbleitertrioden . . . . . *J. Š. Haškovec und A. Klímek*  
 Ein- und ausschaltbare Silizium-Gleichrichter . . . . . *H. F. Storm*  
 Keramische Speicherzellen für außergewöhnliche Umweltbedingungen . . . . . *A. B. Kaufman*

## DIGITALE ELEMENTE

Die Anwendung digitaler Integrieranlagen in der Prozeßführung . . . . . *H. Rechberger*  
 Vorteile und Aussichten der digitalen Drehzahlregelung . . . . . *W. Fritzsche*  
 Regler mit digitalen Rechenoperationen . . . . . *T. M. Aleksandridi, S. N. Diligensky und H. Krug*

## MESSEINRICHTUNGEN FÜR ANLAGEN

Ein universelles Analysiergerät für stochastische Vorgänge . . . . . *J. Krýže*  
 Das Verhalten selbsteinstellender Regler . . . . . *J. L. Douce*  
 Grundlagen und Schaltung eines Vielkanal-Korrelators — Ein Spezialanalogrechner zur statistischen Erfassung regelloser Vorgänge in industriellen Anlagen . . . . . *A. S. Uskov und Yu. M. Orlov*

## ZUVERLÄSSIGKEITSPROBLEME

Die Zuverlässigkeit von Bauelementen — Übersichtsvortrag . . . . . *G. Glinski, B. S. Sotskov und H. Weissmann*  
 Neuartige Steuerventile für redundante elektrohydraulische Systeme . . . . . *K. D. Garnjost und W. J. Theyer*  
 Die Zuverlässigkeit elektrischer Schaltelemente . . . . . *G. W. A. Dummer*  
 Fragen der Betriebssicherheit elektro-mechanischer Übertragungsglieder . . . . . *B. S. Sotskov, I. E. Dekabrun und L. S. Krivorotova*  
 Eine Untersuchung der Zuverlässigkeit von Regeleinrichtungen für Kernreaktoren und Verfahrensregelstrecken  
*L. A. J. Lawrence und R. J. Scotcher*

## SCHLUSSBEMERKUNGEN

*Gy. Boromisza* (Vorsitzender des Technischen Ausschusses im IFAC für Gerätefragen) . . . . .  
*Y. Oshima* (Stellvertretender Vorsitzender des Ausschusses im IFAC für Gerätefragen) . . . . .

# Authors' Names and Addresses

## APPLICATIONS AND COMPONENTS

### AUSTRIA

RECHBERGER, H., Elektrotechnisches Institut, Gusshausstrasse 25, Wien IV

### BELGIUM

DELVAUX, L., Université de Liège, Institut de Mécanique, 75 rue de Val Benoît, Liège

### BULGARIA

MASLAROFF, I. A., The Institute for Scientific Research of Electrical Industry, P.O. Box 67, Sofia 45

### CANADA

TANNER, J. A., Division of Mechanical Engineering, National Research Council, Ottawa 2, Ontario

### CZECHOSLOVAKIA

HAŠKOVEC, J. S., Institute for Information Theory and Automation, Vyšehradská 49, Praha 2

KLÍMEK, A., Institute for Information Theory and Automation, Vyšehradská 49, Praha 2

KRÝŽE, J., Institute for Information Theory and Automation, Vyšehradská 49, Praha 2

ZÁVORKA, J., Institute for Information Theory and Automation, Vyšehradská 49, Praha 2

### FRANCE

CSECH, J., 26 rue de Lyon, Paris 12

HONORÉ, E., Société ANALAC, 101 Blvd. Murat, Paris 16

MONTJEAN, F., 109 rue de Ville d'Avray, Sevres (Seine et Oise)

NASLIN, P., 20 Avenue Pricue de la Côte d'Or, Arcueil (Seine)

RAOULT, J.-C., 16 Avenue de Verdun, Vanves (Seine)

THILLIEZ, J., 175 Blvd. Murat, Paris 16

TURPIN, P., 175 Blvd. Murat, Paris 16

### GERMANY

EHLING, H. J., AEG-Institut für Automation, 1 Berlin N 65, Brunnenstr. 107a

FRITZSCHE, W., AEG-Institut für Automation, 1 Berlin N 65, Brunnenstr. 107a

KINDERMANN, W., Continental Elektroindustrie AG, Askania-Werke, VR-Ta, 1 Berlin-Mariendorf, Grossbeerenstr. 2-10

### HUNGARY

BENEDIKT, O., Budapest Müszaki Egyetem Villamosgépek Tanszék, Budapest XI, Egri József — u. 18

BENEDIKT, S., Institute for Electrical Power Research, Budapest, VI. Rudas L. u. 27

UZSOKI, M., Institute for Electrical Power Research, Budapest, VI. Rudas L. u. 27

VÁMOS, T., Institute for Electrical Power Research, Budapest, VI. Rudas L. u. 27

### ITALY

DE GREGORIO, G., Compagnia Generale di Eletticità, C.G.E., Milano

LITIGIO, G., Instituto Siderurgico Finsider, Genova-Cornigliano

SIRONI, G., Instituto Siderurgico, Finsider, Genova-Cornigliano

STEIN, T., Viale X Giugno 44 B, Vicenza

### JAPAN

ARAKI, K., Institute of Industrial Science, University of Tokyo, Azabu-shinryudocho, Tokyo

FUJII, S., Dept. of Mechanical Engineering, Nagoya University, Furo-cho Chikusa-ku, Nagoya

ISOBE, T., Faculty of Engineering, University of Tokyo, Tokyo

IZAWA, K., Tokyo Institute of Technology, Oh-okayama, Meguro-ku, Tokyo

KANDA, N., Dept. of Mechanical Engineering, Nagoya University, Furo-cho Chikusa-ku, Nagoya

MORINAGA, T., Tokyo Institute of Technology, Oh-okayama, Meguro-ku, Tokyo

NAKANISHI, E., Kyoto University, Kyoto

OSHIMA, Y., Institute of Industrial Science, University of Tokyo, Azabu-shinryudocho, Tokyo

TAKAMATSU, T., Kyoto University, Kyoto

TOTANI, T., Faculty of Engineering, University of Tokyo, Tokyo

### NETHERLANDS

KYLSTRA, F. J., Koninklijke/Shell-Laboratorium, Amsterdam

VAN DER HEYDEN, C.A.J.M., Bataafse Internationale Petroleum Maatschappij N.V.

VAN NES, A. G., Bataafse Internationale Petroleum Maatschappij N.V.

### NORWAY

RUGE, H., Central Institute for Industrial Research, Blindern, Oslo

### SWITZERLAND

BOREL, L., Epul, Lausanne

GLAVITSCH, H., Hertensteinstr. 23, Nussbaumen bei Baden

TSCHUMY, A., 109 rue de Lyon, Geneva

### UNITED KINGDOM

ANDERSON, B. G., Guided Weapons Division, English Electric Aviation Ltd., British Aircraft Corporation Ltd., Stevenage, Herts

DOUCE, J. L., Electrical Engineering Dept., Queen's University (David Keir Bldg.), Belfast

DUMMER, G. W. A., Applied Physics and Technical Services, Royal Radar Establishment, St. Andrews Road, Great Malvern, Worcs.

HERSOM, S. E., Planning and Computer Developments Dept., Richard Thomas & Baldwins Ltd., Spencer Works, Newport Section, Llanwern, Nr. Newport, Mon.

LAWRENCE, L. A. J., Control and Instrumentation Division, Building B. 41, Atomic Energy Establishment, Winfrith, Dorchester, Dorset

MASSEY, R. G., Planning and Computer Developments Dept., Richard Thomas & Baldwins Ltd., Spencer Works, Newport Section, Llanwern, Nr. Newport, Mon.

- M'PHERSON, P. K., Control and Instrumentation Division, Atomic Energy establishment, Winfrith, Dorchester, Dorset
- MUSCETTOLA, M., Control and Instrumentation Division, Atomic Energy Establishment, Winfrith, Dorchester, Dorset
- SCOTCHER, R. J., Control and Instrumentation Div., Building B. 41, Atomic Energy Establishment, Winfrith, Dorchester, Dorset
- TAYLOR, F. J. D., Engineering Dept., G.P.O. Research Station, Dollis Hill, London, N.W. 2.
- TOMLINSON, A., The British Iron and Steel Research Association, 11 Park Lane, London, W. 1.
- WEST, J. C., Electrical Engineering Dept., Queen's University, Belfast
- WHEELER, R. G., 121 Westminster Road, Davyhulme, Urmston, Lancs.
- WISTREICH, J. G., The British Iron and Steel Research Association, 11 Park Lane, London, W. 1.
- U.S.A.**
- BEADLE, R. G., G.E.C., 1 River Road, Schenectady 5, New York
- BEKEY, G. A., Electrical Engineering Dept., University of Southern California, Los Angeles 7, California
- BOOTON, R. C., Jr., Space Technology Lab. Inc., Redonda Beach, California
- BRADFORD, J. T., Jr., Jones & Laughlin Steel Corp., 3 Gateway Center, Pittsburgh 30, Penn.
- CANNON, R. H., Jr., Stanford University, Stanford, California
- EGGENBERGER, M. A., G.E.C., 1 River Road, Schenectady 5, New York
- ELLERT, F. J., General Engineering Lab., G.E.C., 1 River Road, Schenectady 5, N.Y.
- FIEDLER, G. J., Sverdrup & Parcel and Associates, 915 Olive Street, St. Louis 1, Missouri
- GARNJOST, K. D., Moog Servocontrols Inc., East Aurora, New York
- GOODSON, R. E., Purdue University, Lafayette, Indiana
- HOLL, W. H., AC Spark Plug Division, General Motors Corporation, Flint 2, Michigan
- KAUFMAN, A. B., Litton Systems, Woodland Hills, California
- KIRCHMAYER, L. K., Electric Utility Analytical Engineering, G.E.C., 1 River Road, Schenectady 5, New York
- KOCHENBURGER, R. J., University of Connecticut, Storrs, Connecticut
- LANDY, J. J., Sverdrup & Parcel and Associates, 915 Olive Street, St. Louis 1, Missouri
- MERRIAM, C. W., III, General Electric Research Laboratory, G.E.C., 1 River Road, Schenectady 5, New York
- MOCZEK, J. S., Monsanto Chemical Company, 800 North Lindbergh Boulevard, St. Louis 66, Missouri
- NAHI, N., Autonetics, 3311 East La Palma Road, Anaheim, California
- NEWTON, G. C., Jr., Electronic Systems Laboratory, Building 32, M.I.T., Cambridge 39, Mass.
- OLDENBURGER, R., Purdue University, Lafayette, Indiana
- OTTO, R. E., Monsanto Chemical Company, 800 North Lindbergh Boulevard, St. Louis 66, Missouri
- PHILLIPS, R. A., Analytical Engineering, A.P. & P., G.E.C., 1 River Road, Schenectady 5, New York
- RINGLEE, R. J., Electric Utility Analytical Engineering, G.E.C., 1 River Road, Schenectady 5, New York
- ROBERSON, R. E., 1100 No. Cerritos Drive, Fullerton, California
- SMYTH, R. K., Autonetics, 3311 East La Palma Road, Anaheim, California
- SOLLECITO, W. E., Electronics Park, G.E.C., Syracuse, New York
- STORM, H. F., G.E.C., 1 River Road, Bldg. 37-575, Schenectady, New York
- THAYER, W. J., Moog Servocontrols Inc., East Aurora, New York
- TROUTMAN, P. H., G.E.C., 1 River Road, Schenectady 5, New York
- WILCOX, R. B., R.C.A. Aerospace Division, Burlington, Mass.
- WILLIAMS, T. J., Monsanto Chemical Company, 800 North Lindbergh Boulevard, St. Louis 66, Missouri
- YOUNG, L. R., Room 32-101, Massachusetts Institute of Technology, Cambridge 39, Mass.
- U.S.S.R.**
- At the Institute of Automation and Telemechanics, Moscow I-53, Kalachevskaya 15-A:
- ALEKSANDRID, T. A.
- DOMANITSKY, S. M.
- KHOKHLOV, V.
- LERNER, A. Ya.
- SOTSKOV, B. S.
- TAL, A. A.
- At the U.S.S.R. National Committee on Automatic Control, Moscow I-53, Kalachevskaya 15-A:
- ANISIMOV, I. V.
- USKOV, A. S.
- VASILYEV, Yu. K.
- VOLTER, B. V.
- KOROBKO, M. I., Kiev 52, Nagornaya 22, Institute of Automation, Gosplan Ukrainian S.S.R.
- KRUG, G. K., Moscow E-250, Krasnozarmennaya 14, Moscow Power Institute
- VENIKOV, V. A., Moscow E-250, Krasnozarmennaya 14, Moscow Power Institute
- YUGOSLAVIA**
- TOMOVIĆ, R., Beograd, Dobračina 13
- The following authors have contributed to the volume on **THEORY**

## THEORY

### ARGENTINE

ROXIN, E., Directorio 1304, Haedo (FNDPS), Prov. B. Aires

### CANADA

SMITH, H. W., 74 Reid Avenue, Ottawa 3, Ontario

### CHINA

HANG KING-CHING, Institute of Mathematics, Academia Sinica, Peking

HWANG LING, Peking University, Dept. of Mathematics, Peking

SUN JIAN, Institute of Mathematics, Academia Sinica, Peking

### CZECHOSLOVAKIA

BENEŠ, J., Institute for Information Theory and Automation, Vyšehradská 49, Praha 2

PETERKA, V., Institute for Information Theory and Automation, Vyšehradská 49, Praha 2

STREJC, V., Institute for Information Theory and Automation, Vyšehradská 49, Praha 2

### FRANCE

GILLE, J.-C., Ecole Nationale Supérieure de l'Aéronautique, 32 Blvd. Victor, Paris 15

GUIGNABODET, J., 55 rue Caulaincourt, Paris 18

LEFEVRE, P. M., 3 rue Claude Matrat, Issy les Moulineaux (Seine)



LOEB, J., 14 rue Alphonse Moguez, St. Cloud (Seine et Oise)  
PÉLÉGRIN, M. J., 11 bis rue de la Planche, Paris 7

#### GERMANY

LAUBER, R., AEG-Institut für Automation, 6 Frankfurt/M.-Hausen, Tilsiter Str. 5  
LEONHARD, A., Lehrstuhl für elektrische Anlagen an der Technischen Hochschule Stuttgart, 7 Stuttgart N, Breitscheidstr. 3  
MESCH, F., im Institut für Regelungstechnik der Technischen Hochschule Darmstadt, 61 Darmstadt, Schlossgraben 1  
PAVLIK, E., 75 Karlsruhe, Lassallestr. 9

#### INDIA

PRASAD, T., Bihar Institute of Technology, P.O. Sindri Institute, Dhanbad (Bihar)  
SINHA, V. P., Bihar Institute of Technology, P. O. Sindri Institute, Dhanbad (Bihar)

#### ISRAEL

BONENN, Z., Scientific Dept., Israel Ministry of Defence, P.O.B. 7063, Hakirya, Tel-Aviv

#### ITALY

MARACCI, R., Montecatini SPEB/PROS, Largo Donegani 1, 2, Milano

#### JAPAN

HAYASHI, C., Dept. of Electrical Engineering, Kyoto University, Kyoto  
IWAI, S., Dept. of Electronic Engineering, Kyoto University, Kyoto  
KONDO, B., Dept. of Electronic Engineering, Kyoto University, Kyoto  
SAKAWA, Y., Dept. of Electrical Engineering, Kyoto University, Kyoto

#### NETHERLANDS

VAN DEN BOS, A., Dept. for Technical Physics, Technological University of Delft, Delft  
VAN DER GRINTEN, P.M.E.M., Staatsmijnen in Limburg, Central Laboratory, Geleen  
VELTMAN, B.P.TH., Dept. for Technical Physics, Technological University of Delft, Delft

#### NORWAY

BALCHEN, J. G., Eidanger Salpeterfabriker, Herøya, pr. Porsgrunn  
BLANDHOL, E., Eidanger Salpeterfabriker, Herøya, pr. Porsgrunn

#### POLAND

JAROMINEK, W., Warszawa ul Grójecka 40a m 34  
KULIKOWSKI, R., Polski Komitet Pomiarow I Automatyki, Warszawa ul Czachiego 3/5  
PASZKOWSKI, St., Polski Komitet Pomiarow I Automatyki, Warszawa ul Czachiego 3/5  
STRASZAK, A., Polski Komitet Pomiarow I Automatyki, Warszawa ul Czachiego 3/5

#### ROUMANIA

MOISIL, Gr. C., Academia R.P.R. Institutul de Matematica, Str. M. Eminescu 47, Bucuresti 3

#### SWEDEN

PERSSON, E. V., Research Laboratory, A.S.E.A., Västerås  
QVARNSTRÖM, B., Chalmers Institute of Technology, Gibraltargatan 5 R, Gothenburg

#### SWITZERLAND

ELDIN, H. NOUR, Swiss Institute of Technology, Gloriastr., Zurich 6  
PERRET, R., Division Internationale, Battelle Memorial Institute, 7 route de Drize, Geneva  
ROUXEL, R., Division Internationale, Battelle Memorial Institute, 7 rout de Drize, Geneva

#### UNITED KINGDOM

ADEY, A. J., Dept. of Engineering, University of Cambridge, Cambridge  
ANDREAE, J. H., Senior Engineer, Standard Telecommunication Laboratories Ltd., London Road, Harlow, Essex  
BROWN, B. M., Royal Naval College, Greenwich, London, S.E. 10.  
COALES, J. F., Dept. of Engineering, University of Cambridge, Cambridge  
FARMER, E. D., Central Electricity Research Laboratories, Cleve Road, Leatherhead, Surrey  
KERR, C. N., Engineering Laboratories, The University, Glasgow, W. 2.  
MACLELLAN, G. D. S., Engineering Laboratories, The University, Glasgow, W. 2.  
MERKLINGER, K. J., Engineering Laboratory, University of Cambridge, Trumpington Street, Cambridge  
ROSENBROCK, H. H., University of Cambridge, Dept. of Engineering, Engineering Lab., Trumpington Street, Cambridge  
STILES, J. A., Dept. of Engineering, University of Cambridge, Cambridge

TAYLOR, W. K., Dept. of Electrical Engineering, University College, Gower St. London, W.C. 1.  
WESTCOTT, J. H., Imperial College, Exhibition Road, London, S.W. 7.

#### U.S.A.

CHANG, S.S.L., Dept. of Electrical Engineering, New York University, University Heights, New York 53  
DESOER, C. A., Dept. of Electrical Engineering, University of California, Berkeley 4, California.  
EVELEIGH, V. W., Electronics Laboratory, G.E.C., Electronics Park, Syracuse, N.  
FLAKE, R., Electrical Engineering Dept., Washington University, St. Louis 30, Missouri  
FLÜGGE-LOTZ, I., Division of Engineering Mechanics, Stanford University, Stanford, California  
FRIEDLAND, B., General Precision Inc., Aerospace Research Center, 1150 McBride Avenue, Little Falls, New Jersey  
GAYLORD, R., Aerospace Corporation, P.O. Box 95085, Los Angeles 45, California  
GIBSON, J. E., Control and Information Systems Laboratory, School of Electrical Engineering, Purdue University, Lafayette, Indiana  
HORING, S., Bell Telephone Laboratories, Whippany, New Jersey  
JURY, E. I., Dept. of Electrical Engineering, University of California, Berkeley 4, California  
KELLEY, H. J., Grumman Aircraft Engineering, Corp., Bethpage, Long Island, New York  
LASALLE, J. P., RIAS, 7212 Bellona Avenue, Baltimore 12, Maryland  
LEE, E. B., Minneapolis Honeywell Regulator Co., 2600 Ridgway Road, Minneapolis, Minnesota  
LI, C. C., Dept. of Electrical Engineering, University of Pittsburgh, Pittsburgh Pennsylvania  
MEDITCH, J. S., Aerospace Corporation, P.O. Box 95085, Los Angeles 45, California  
NESBIT, R. A., Aerospace Corporation, P.O. Box 95085, Los Angeles 45, California  
NEUSTADT, L., Aerospace Corporation, P.O. Box 95085, Los Angeles 45, California  
SARACHIK, P. E., Dept. of Electrical Engineering, Columbia University, New York 27.  
SZEGÖ, G. P., Control and Information Systems Laboratory, School of Electrical Engineering, Purdue University, Lafayette, Indiana  
TITUS, H. A., Jr., U.S. Naval Postgraduate School, Monterey, California

# AUTHORS' NAMES AND ADDRESSES

## U.S.S.R.

At the Institute of Automation and Telemechanics, Moscow I-53, Kalachevskaya 15-A:

AIZERMAN, M. A.

ALIMOV, Yu. I.

BUTKOVSKII, A. G.

FELDBAUM, A. A.

GAVRILOV, M. A.

GULKO, F. B.

KASAKOV, I. E.

KRASOVSKII, A. A.

KRAZOVSKII, N. N.

MEEROV, M. V.

NOVOSELTSEV, V. N.

PETROV, B. N.

POSPELOV, G. S.

TSYPKIN, Ya. Z.

At the U.S.S.R. National Committee on Automatic Control, Moscow I-53, Kalachevskaya 15-A:

ANDREEV, N. I.

BARBASHIN, Ye. A.

KLUBNIKIN, P. F.

POPOV, E. P.

STRATONOVICH, R. L.

VOLGIN, L. N.

CHICHINADZE, V. K., Tbilisi 42, Pekin 56, Institute of Electronics, Automation and Telemechanics, Academy of Sciences, Georgian S.S.R.

KUNTSEVICH, V. M., Kiev 54, Chkalova 55-6, Institute of Electromechanics, Academy of Sciences, Ukrainian, S.S.R.

VORONOV, A. A., Institute of Electromechanics, Leningrad D-41, Dvortsovaya Naberezhnaya 18

# The Second International Congress of I.F.A.C. in Basle 1963

EDUARD GERECKE, Third President of I.F.A.C.

This, the second I.F.A.C. Congress, is being held in Basle from 27th August to 4th September. 1476 participants and 200 ladies are present from the following countries: Argentine (1), Austria (11), Belgium (37), Bulgaria (8), Canada (5), China (10), Congo (3), Czechoslovakia (11), Denmark (10), Finland (25), France (173), Germany (212), Hungary (30), India (1), Israel (4), Italy (47), Japan (19), Mexico (1), Netherlands (83), Norway (22), Poland (32), Portugal (3), Roumania (10), Spain (6), Sweden (44), Switzerland (237), Turkey (7), United Kingdom (160), United States of America (154), Union of Socialist Soviet Republics (80), Yugoslavia (30).

As Automatic Control today covers a very large field, the Executive Council of I.F.A.C. selected the following limited fields for the Congress.

1. *Theory*
2. *Applications*
3. *Components*

As an innovation, 11 Survey Papers dealing with the actual state of automatic control in most fields of theory, applications and components will be read. All Congress delegates can attend these lectures, and will, I am sure, certainly appreciate the opportunity of getting a competent survey from outstanding specialists of today's position and of future developments in the different fields of Automatic Control.

Two hundred and sixty Discussion Papers were submitted to

I.F.A.C. in September 1962, and the I.F.A.C. Selection Committee accepted 159 of these, namely 82 on Theory, 57 on Applications, and 20 on Components. Fifty half-day sessions are planned for the discussion of these papers, 25 of them on Theory, 19 on Applications and 6 on Components. Undoubtedly the I.F.A.C. Basle Meeting will contribute to the promulgation of the new and more advanced chapters of Automatic Control.

The organization of such a large meeting requires the co-operation of a large number of individuals. The I.F.A.C. officials and many members of the Technical Committee did a great deal of preparatory work, all on a voluntary basis, and their help is very much appreciated. Our sincere thanks also go to the Honorary Secretary, Dr. G. Ruppel, Düsseldorf, for his indefatigable secretarial work and to the Honorary Editor, Professor V. Broïda, who prepares the Discussion Papers and the Proceedings with the aid of the Co-editors, D. H. Barlow, London, and Professor O. Schäfer, Aachen. I should like also to thank the Scientific Secretary, E. Ruosch, Zurich, for his most valuable assistance.

The expenses of the Congress, including the Preprints and the Survey Papers, amount to U.S. \$ 97,000 for which the Council members of the Swiss Society of Automatic Control are personally responsible. As third President of I.F.A.C. I should like to thank most heartily the Swiss Federation of Automatic Control and all those who contribute financially, especially the Authorities of the Canton of Basle Town, the Swiss Industries Fair and the Swiss Industry.

# How I.F.A.C. was Founded

H. CHESTNUT, First President of I.F.A.C.

It is appropriate as we start this Second Congress of I.F.A.C. on Automatic Control that we look back at some earlier international meetings on automatic control and see what we may learn from the past that will give us some ideas that will be useful for the future. Within the past twenty-five years the science and art of automatic control has grown rapidly and interest has developed in such subjects as regulation, automation, and automatic control on the part of engineers, politicians, and laymen. We who are developing and using the techniques of automatic control are, in a way, charged with the responsibility of making automatic control most effective. It is important that the place of international activities, including meetings and congresses, be considered as a part of the process of helping us to accomplish our job of making automatic control better.

Although the first General Assembly of I.F.A.C. was held in Paris in September 1957, a number of meetings that were international in character if not in name had been held prior to that time. In 1951 at Cranfield, England, was held an outstanding meeting on 'Automatic and Manual Control' sponsored by the Department of Scientific and Industrial Research, the Proceedings of which were edited by Professor Arnold Tustin. In 1953 in New York under A.S.M.E. sponsorship was held a 'Frequency Response Symposium' that was international in its list of authors; its Proceedings were edited by Dr. Rufus Oldenburger.

By 1956 the heightened interest in automatic control caused a number of meetings to be held in Europe on or related to this subject. Outstanding among these was 'Regelungstechnik' in Heidelberg, Germany, in September 1956.

An earnest effort was made to obtain broad international participation on the part of authors from many countries serving to bring together many leading men in automatic control. This had the effect of increasing their interests in the possibilities of more regular and more internationally organized meetings on automatic control theory and practice that would be more broadly publicized and attended.

It was appreciated that similar progress in automatic control was being made in many parts of the world, in many cases with each group of people largely oblivious of the works of the others or without even knowing who these other people were. Furthermore, it was apparent that a great amount of time and effort would be required to push forward our understanding of automatic control and our ability to use it. Hence the help and benefit to be obtained by the sharing of ideas and information on an international basis would be valuable in making for more rapid progress.

Fortunately, a group of dedicated men, including Dr. Victor Broida, Chairman; Dr. Otto Grebe, Professor A. M. Letov,

Professor P. J. Nowacki, Dr. Rufus Oldenburger, Dr. G. Ruppel, and Mr. D. B. Welbourn, were willing to devote their energies during the period September 1956 to August 1957 to helping in bringing about the formation of the International Federation of Automatic Control at the first General Assembly of I.F.A.C. in Paris, September 1957.

At this initial meeting a Constitution was adopted built around the principle of national rather than personal representation, and particular emphasis was placed on the holding of periodic international congresses. Another important objective was 'to promote the science of automatic control by the interchange and circulation of information on automatic control activities in co-operation with national and other international organizations'. Plans were laid for the first I.F.A.C. Congress to be held in Moscow, U.S.S.R., in 1960 about which you will hear shortly from Professor A. M. Letov who was President of I.F.A.C. at that time. The first I.F.A.C. Congress was a very successful one from which was obtained a fine, four volume set of Proceedings edited by John F. Coales.

The International Federation of Automatic Control is indeed fortunate that the Swiss Federation of Automatic Control was willing to serve as host for this second I.F.A.C. Congress here at Basle. The preparation and labour involved by members of the Swiss Federation of Automatic Control for this Congress are outstanding. On behalf of all those present here, I want to thank Professor Gerecke in his dual role as President of I.F.A.C. and as President of the Swiss Federation of Automatic Control for his help and for the efforts and skill of the members of the two groups which he represents and directs.

As first President of I.F.A.C., I am very pleased with the way that I.F.A.C. has developed and with the interest which its Congresses have stimulated. The growth of I.F.A.C. symposia on special subjects has been of particular interest to me. I look forward with hope for additional symposia on suitable topics of concern to experts in the field of automatic control.

We all have reason to be pleased with the fine progress that has been made in our automatic control accomplishments to date. Lest we become complacent, however, I should like to point out two other fields for serious attention by control people. These are:

- (1) The need for 'optimizing the process of making automatic control', i.e. bridging the gap between theory and practice.
- (2) The need for working with qualified people in the social, economic, and political fields to help make the net effect of automatic control and automation a cause for hope rather than a reason for fear.

For the past 5 years and more, one of the most popular topics of automatic control investigation has been the subject of optimum control. With all this emphasis on optimum control, it would appear possible to apply some of these general principles to permit us to perform the process of designing and building automatic controls more quickly, more cheaply, more reliably, or more favourably in some other sense which allows the desired balance of a number of these objectives. Not only should concern be given to determining how to make a system which will be optimum, but also the process of design itself should be one which can be readily applied by the large number of designers who will be applying these ideas to the making of better controls. Attention must be given from the theory point of view to including the practical application of these new control concepts. Efforts must be made from both sides to bridge the gap between theory and application.

The second problem which deserves more of our attention is that of automation, the popular term by which much of automatic control is known. Throughout the centuries, Man has hoped to find a way of obtaining goods and services with a minimum of effort to himself. With the advent of more and more automation, we are approaching the condition where a signifi-

cant proportion of the necessary production and services can be achieved with a minimum amount of direct human effort.

To realize the opportunity that automation can afford will require more than just the technical attention that we, working in automatic control, can apply. The changes in production and services brought about by automation will also involve changes in the way people live and earn their livelihood. I believe it will be more effective if engineers and scientists skilled in automatic control systems work with people skilled in dealing with social, economic and political problems to help bring about the needed changes in a smooth and socially acceptable fashion.

Unfortunately, the human time constant is one of the longest we have to deal with. Although the problems associated with the introduction of more widespread automation are great. The opportunities for a better world at peace make the challenge of using automation for the betterment of man one that is certainly worth working for.

From what I have seen of the preprints of the papers for this Congress, I am looking forward to an interesting week of discussions here. I am hopeful that the creativity and vision that have characterized I.F.A.C. from its beginning will continue to grow and flourish as we move ahead with this second I.F.A.C. Congress and on to the future.

## The First I.F.A.C. Congress

A. M. LETOV, Second President of I.F.A.C.

Three years have passed since the Moscow Congress—the first I.F.A.C. Congress, which brought together 2,000 specialists from 28 countries; and at which some 300 papers dealing with the solution of major scientific and engineering problems of automation, were read and discussed.

Although reports of the Congress were published in newspapers and magazines in many countries, and the Congress Proceedings were published in Russian and English, I now remember the Congress not merely because it brought great satisfaction to those who took part, or because the aims of I.F.A.C.—so well expressed in the speech of our President, Professor Gerecke—were realized so widely for the first time in Moscow.

I recall the Congress for another reason; because, as the poet said, 'pleasant recollections are the fount of good inspiration', from which we draw new strength to develop the future activity of our Federation. The conditions for developing this activity are becoming more and more favourable.

We have now assembled at our second Congress. It has been organized by the Swiss National Federation, headed by our respected President Professor Gerecke. But the activity of the Federation is not characterized by this alone. The activity of I.F.A.C. is characterized, in particular, by the willingness of many other countries to organize subsequent I.F.A.C. Congresses. Such Congresses will undoubtedly be held.

Although it may be in 90 years time, I still hope fervently

that I can live to the noteworthy day when the I.F.A.C. Congress will have gone round all the countries in our Federation and returned once again to Russia—perhaps to Moscow—involving not 2,000 but 20,000 participants. I also look forward to the day when the linguistic difficulties of communication will have been overcome by the creation of miniature 'radio-computers' which will translate into one's own language the speeches of representatives of all countries on the globe; and to the day when science and technology will make it both pleasant and inspiring to look forward to what is to come a century ahead.

You will say that it is a very remote dream. As yet it is a dream, I grant you. Let me just say this. First, you are all people who do creative work—dreamers—and all the plans you implement so wonderfully begun with a dream.

Secondly, let me, by speaking of my dreams, give those with a sense of humour a chance to say, what else can a man do who, after 6 years of helping to run the Federation, and for the moment still its Past President, but relinquish the authority in three days when he retires.

Thank you, Mr. President, for the opportunity given me here to dream aloud; to my audience I say thank you for your attention.

With all my heart I say, 'I look forward to meeting soon at a new Moscow Congress, dear colleagues'.

# The Information Revolution and Its Impact on Automatic Control

I. L. AUERBACH, President of the IFIP

The invention of the electronic digital computer in 1946 marked the beginning of the information revolution. The ensuing seventeen years have seen a development very much like the industrial revolution that followed the advent of the steam engine in 1765. In both instances, the advancement of civilization had created a growing need for new ways of accomplishing vital tasks. The major technological breakthrough not only filled the need, and opened new avenues in many fields, but led to the discovery of new fields, where development would have been out of the question without the new tool. Thus it was that the electronic computer brought about a revolution in information processing, rather than just advanced normal development in the field.

The essential feature of the industrial revolution was man's amplification of his brawn by the use of engines. In the information revolution the emphasis has shifted to the amplification of brain through computers and information processing systems. Already there are as many different kinds of computers as there are kinds of engines. The applications of digital data processing are limited only by the ingenuity of scientists and technologists in their respective technical fields.

There is a basic and fundamental difference between brawn and brain—a difference that is exaggerated when these faculties are extended and amplified by mechanical engines or electronic processors. A muscle or an engine consumes energy to accomplish work. The fuel can never be recovered or re-used. In contrast, information handling is non-destructive. Information is used and applied without being consumed. It can be used over and over again, and many of the applications augment the original supply, but none can diminish it.

The words in a book, for example, remain intact, and can be read by any number of people without loss of information content. The responses the words elicit in different minds may even go beyond the original content. Each new edition of a great book may have more footnotes than the one before it—words growing on words as they strive to capture ideas. Similarly, a computer manipulates information in a scientific computation, and generates new information from it. The information is never consumed by being processed; it can be retrieved and used repeatedly in many different ways.

This difference perhaps explains why the information revolution has been faster and more widespread in its implications than the industrial revolution. Electronic data processing is so broad a subject that only an infinitesimal fraction of it can be explored in a short time. I would like to take a brief look at the field where the engine and the computer work together; the impact of the information revolution on automatic control. In addition, at the request of your President, I will review the background of the International Federation for Information Processing (IFIP) and discuss the areas of cooperation between that body and the International Federation for Automatic Control (IFAC).

Figure 1 shows the basic structure of an automatic control loop. In it the status of a physical system is detected through

sensors or transducers and transmitted to a device labeled 'Computer'. This device identifies any differences between the indicated status of the system and its desired status, and activates controls to modify the physical system. The changes in status resulting from this modification are in turn sensed and transmitted to the computer, and so the operation of the loop goes on. The computer not only can guide the system through a series of steps, but can modify any future step on the basis of the results of previous steps.

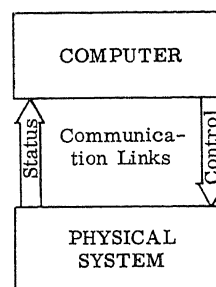


Figure 1. The automatic control loop

This simplified diagram illustrates how the computer performs the functions of a human operator. With its fantastic speed of reaction and calculation, its resistance to fatigue or distraction, and the great variety of inputs and outputs it can utilize, the computer can replace a human operator in many physical systems. More important, it can assume new tasks too taxing for human beings and thus make new physical systems feasible. The launching of a satellite, for example, simply could not be handled by anything short of an automatic computer system.

Still man's place in automatic control is a vital one. The loop shown in Figure 1 is usually part of a more complex loop as shown in Figure 2. The 'desired status' mentioned in the discussion of Figure 1 is determined by man. He observes the physical system, and applies his intelligence, his ability to judge situations rather than respond to them deterministically, and his set of values. He establishes the criteria and the end goals to be achieved by the system, and communicates his conclusions to the computer via programming. This is a critical input to the computer. It cannot be absent, as might be inferred from Figure 1, although it may precede in time the actual operation of the system.

This powerful combination of man, computer, and physical system multiplies the resources of all three. Automatic control combines the benefits of the industrial revolution and the information revolution, giving man an extension of his brain and his brawn, both applied to the same task. The digital computer makes this system quite distinct from the man-engine combination alone. It handles great quantities of data at tremendous speeds, extending the realm of operations that can be performed. Programming techniques also combine decision making with mathematical calculations, and this permits dis-

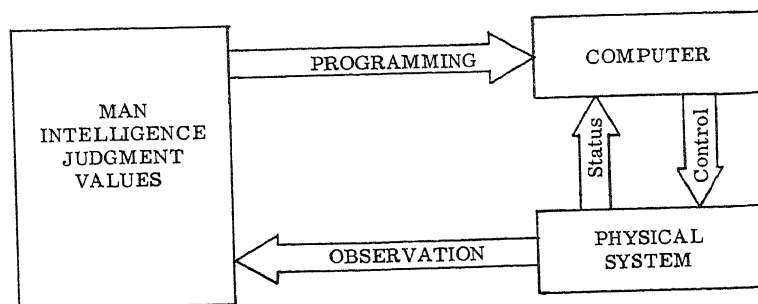


Figure 2. The man loop

continuous or discrete control to be intermixed with the various kinds of continuous or smooth control that are afforded by analogue devices.

It can be seen in Figure 2 that man's portion of the loop, proceeding from the physical system to the computer, parallels the direct sensing of system status by the computer. Both connections are necessary, as each exploits the individual characteristics of its medium: man or machine. Some of the outstanding developments in recent computer history are directed toward man's turning over as many as possible of his own functions to the computer. Much of the programming process is being made automatic, and can be classified as computer reaction to a system status.

The computer may be multi-programmed, that is, may time-share or interleave a number of independent or interrelated programmes run at the same time. It is now feasible to have one computer run a programme that controls the physical system based on the results of an independent programme being run at the same time that derives economic and business criteria. For example, if the inventory and order control for a business were maintained by the same computer, then the output of the physical system would be based on the current requirements and permit very rapid response times. In addition, multi-computer systems are becoming more commonplace, and with appropriate communication links further broaden the variety of tasks that can be handled by computers in the automatic control loop.

The programming of the computer in the automatic control loop is far more complex than most people anticipated. The programme is actually interposed between the computer and the process or physical system it is to control. It is a necessary link in the computer controlled system. In this area, there is need for more work in automatic programming techniques and compilers specifically oriented toward process control. Multi-programme and multi-computer systems will add to the complexity and impose the need for far more sophisticated interrupt features and the techniques for their application. The skill required to formulate the computer programme is major and is far more difficult in complex interconnected computer systems. It should be clear that greater training for people capable of doing this work is essential.

It must be recognized that more and more is being done by computers that it was thought only human beings could do. Pattern recognition is one such extension of computer capabilities. There exist today devices for recognizing printed symbols in a number of different type fonts without human assistance. This enables the introduction of text from a printed page into a computer, where formerly translation into a digital form such as holes in paper was necessary. Similarly, it is possible for

computers to read maps and match them with the areas they represent.

Rather than belittle the activities of man's brain, these imitations of thought by computers show man's superior intelligence more clearly. The programming of artificial perception calls for not only the same perception on man's part, but a transcendent awareness of how that perception works. By reflecting on the way his own mind operates, man has been able to reduce many of its functions to the sort of simple instructions a computer can follow.

This has not been an easy process. In language translation, for example, great difficulties were encountered. Linguists were not immediately aware of how they translate. But linguists and computer engineers are consulting and coming to grips with the problem, and teaching both participants more about language than was ever known before. Completely new methods of linguistic analysis are opening up, which will be as useful to future language study itself as they are to computer applications.

Artificial learning is another field that has come into being as a result of mechanizing thought processes. The computer in the three-way loop can modify its own programme according to occurrences in the physical system. A man can programme a computer to play a game and then be beaten by the computer. Investigations in this area are shedding new light on the problems of educational psychology.

The area in which computer technology can be expected to make the most significant contribution to automatic control is in the optimization of continuous processes. This optimization will be achieved in two ways. The computer's computational capabilities will be utilized to make more effective use of data on the internal variables of the process. The computer also has the completely unique ability to help man to optimize a process on the basis of variables external, but related, to the process.

Process optimization on the basis of internal variables is coming about through the development of computer programmes that continually modify process parameters in the light of empirical experience. These are really automatic learning programmes in which the computer uses a model of the process to predict the response of a change in an operating variable. The computer then makes the change in the variable, observes the response (through sensory instrumentation), and modifies predictive methodology in the light of its empirical experience.

Process optimization on the basis of external variables is a concept that depends entirely upon the computer's unique capabilities. Going well beyond the concept of integrating the various loops in an individual process, work in this area is directed at integrating several individual but related processes,

plus such business variables as product orders, delivery, and inventory.

The benefits that can result from these two approaches to process optimization are obvious. However, the problems of realizing these benefits are extremely formidable. Success hinges upon how well we can integrate man, machine and the process into an effective closed loop. To do this requires an interdisciplinary effort on the part of the control, operating, and computer engineers.

Much work must be done on control theory and on formalizing the operating methods that are still largely intuitive. Before we can programme computers to effectively perform the process-control functions, we must deepen our understanding of the basics of continuous processes, and we must be able to establish explicit criteria for defining an 'optimum' condition. We must also learn how to translate the 'feel' human operators have for the process into computer programmes.

The effective application of the digital computer to process control is not only an analytical problem; it is also a software problem of major proportions. After the analyses, it is the computer systems engineers who must accept the final responsibility for developing and applying the proper software and programming skills, such as multi-programming and multi-computer programming, that will effectively relate the computer, the processes and the man to each other.

In a limited way, computers are reciprocating the effort to make them behave like men. They seem to be showing men how to behave like computers. The quantities of information handled in modern data processing systems have necessitated new approaches to the problems of indexing, storing, and retrieving information. The newly developed methods can be applied to situations not involving electronic computation at all. Even if all the electronic computers in the world ceased to exist, the effects of the information revolution would still be felt, and would exercise their influence on the future of civilization.

The potentialities of this revolution can best be realized with maximum communication within the field. This interchange of ideas should not be inhibited by national boundaries. It calls for international cooperation and international organizations such as IFIP and IFAC. Now I will turn to a discussion of the international Federation for Information Processing, the first international organization dedicated to all facets of the information processing sciences. As defined by IFIP, these information processing sciences include theory, mathematics, equipment, and application—all applied to the collection, transmission, computation, translation, storage, retrieval, reduction, and display of information.

As stated in its statutes, the aims of IFIP are threefold:

- (a) Sponsor international conferences and symposia on information processing, including mathematical and engineering aspects.
- (b) Establish international committees to undertake special tasks falling within the spheres of action of the member societies.
- (c) Advance the interests of member societies in international cooperation in the field of information processing.

One of the goals of IFIP is to expose the people of the world (those who will be affected by information technology as well as those directly associated with it) to some idea of the progress that can be made through the intelligent use of the electronic digital computer. We hope to make an increasingly greater

number of people aware of the information processing sciences and the benefits that can be derived from them.

In achieving the above aims, IFIP fulfills the need for better worldwide communication and increased understanding among scientists of all nations of the role information processing can play in accelerating technical and scientific progress. It is hoped that better dissemination and centralized control of information about computer technology and application techniques will result in greater scientific advances, which will achieve the two purposes of benefiting mankind and advancing the state of the art.

The first International Conference on Information Processing, sponsored by UNESCO, in June, 1959, provided a forum for the meeting of 1800 delegates from 37 countries. During the planning for this conference, it became apparent that future international meetings and other activities were essential to the development of the information sciences in many countries of the world.

On June 18, 1959, recognizing the importance and success of the UNESCO conference, representatives of the computer societies of 18 countries met in Paris to formulate the preliminary structure of IFIP. Statutes for the federation were drafted, and upon the agreement of 13 national technical societies (six more than the minimum required), IFIP came into official existence in January, 1960.

#### *IFIP Congress 62*

The IFIP Congress 62 was attended by more than 2800 scientists from 41 nations, who were exposed to a comprehensive survey of the technical achievements and goals that have been made possible and practical by the digital computer. This experience demonstrated to them the profound effect that the information revolution is having upon mankind.

In addition to the technical sessions at the Congress, the computer scientists and users attending had an opportunity to view an exhibition of the progress being made in hardware development. The INTERDATA exhibition included exhibits of 48 companies from eight nations. This exhibition, with its emphasis upon hardware, complemented the material presented in the technical sessions on the software and application aspects of information processing.

For the practical realization of its goals, IFIP has established three technical committees and one working group. The scope and accomplishments of these technical committees are summarized below.

#### *IFIP TC-1 Terminology*

The scope of this committee is the establishment of terminology of digital computers and data processing devices, equipments, media and systems. The objective is to promote the exchange of information, leading to the compilation of a multi-lingual glossary for information processing systems and related subjects.

The specific programme of work of this committee includes: select natural languages for which the terms shall be defined; collect documentation of pertinent glossaries; systematize a master list of terms and concepts requiring definition; create or adapt terms for missing concepts and assign these terms to one or more specific fields; choose, modify or originate accepted definitions.



To avoid duplication of effort, the IFIP Committee on Terminology was affiliated with a similar committee within the International Computation Centre of Rome to form 'IFIP/ICC TC-1 Terminology'.

This committee has met seven times to date, and by the end of 1963 will have defined and assigned words in the English language to approximately 1000 concepts. Then specific language area groups will translate these concepts into the major national languages so that the particular words may be chosen in whatever language describes each of the concepts most accurately. It is expected that a vocabulary of 1000 concepts will be published during 1964.

It is significant that the multilingual glossary being developed by IFIP/ICC TC-1 Terminology has been officially requested by the International Standards Organization through subcommittee 1 of its technical committee 97. ISO will use this glossary as the basic input in their effort to establish an international standard.

#### *IFIP TC-2 Programming Languages*

The scope of this committee is to promote the development, specification, and refinement of common programming languages with provisions for future revision, expansion and improvement. This specific programme of work includes: general questions on formal languages such as concepts, descriptions and classifications; the study of specific programming languages; and the study and, where appropriate, coordination of the task of developing new programming languages for which there appears to be a need.

#### *Working Group 2.1 ALGOL*

During the past year, a series of meetings was held between the original ALGOL authors and other experts who clarified and amplified certain aspects of ALGOL '60 language, removing ambiguities that existed. Through IFIP TC-2 the working group submitted the ALGOL '60 document to the council of IFIP which approved the document and made ALGOL '60 official IFIP language. Once again the International Standards Organization, through Subcommittee 5 of Technical Committee 97, has requested that certain specific additions to the ALGOL

language be considered and then submitted to ISO for consideration as an international standard.

#### *IFIP TC-3 Education*

At the August 1962 meeting of the Council of IFIP it was decided to form a new committee on education. The objective of this committee is to establish comprehensive training programmes and suggested curricula for the education of technical people from all over the world who are in fields in which the computer can make a significant contribution. Another function of the committee will be to generate material to acquaint the lay public with the computer and its impact on the various aspects of society. This committee will, in fact, serve as a central clearing house on all educational material pertaining to the information processing sciences. In this capacity, it will assist in preparing or providing translations, lists of available material, and other necessary information services.

The membership of each of the above IFIP committees is international, assuring each national group the opportunity to review and comment on all committee work before it reaches final, rigid form. This policy provides the committees with the additional advantage of having a consensus that includes the viewpoints of many diverse backgrounds.

The activities of IFIP require the time, effort and interest of many people from many parts of the world, all of whom are devoting their services willingly and without material reward. They are to be commended on their accomplishments, for through their endeavours they are demonstrating that people from diverse national backgrounds can work together and communicate effectively to achieve a worthwhile international goal.

There are many areas in which the work of IFIP and that of IFAC impinge upon each other. Cooperation can take place through the common membership of the same scientists in the national technical societies who are members of both our respective federations and by more direct cooperation of our technical committees. It is hoped that ways will be found in the coming years to increase the cooperation and coordination of the activities of these two international federations, and so to provide means for each of the federations to achieve its own goals more effectively and more quickly.



# THEORY



# NON-LINEAR SYSTEM THEORY

---

## Statistical Methods in Automatic Control

A Survey by V. S. PUGACHEV

### Introduction

Every automatic system performs its function, subject to the permanent effect of various noises in its elements, as well as external interferences, natural or artificial, sometimes specially organized. Hence statistical methods are especially important for modern automatic control theory.

Statistical methods afford the evaluation of automatic systems accuracy, as well as the determining of optimal system characteristics, which provide maximum possible accuracy for a given statistical behaviour of signals and noises. Without statistical methods it is impossible at present to design complicated systems. For example, the development of modern theory of self-adaptive systems is possible only on the basis of statistical methods. Only statistical methods afford the determining of dynamic characteristics of plants and control systems from normal operating records with real inputs (more precisely the designing of the best, in a sense of a certain criterion, models for given systems, because the dynamic characteristics of any system are subject to continuous random variations, and no absolutely precise measurements are available).

Statistical methods are necessary not only for the design of control systems, but in many cases for organizing control processes, i.e. for the shaping of control signals in the systems elements. This is especially important for complicated systems containing many elements and communication channels transmitting information among the elements of a system. Control signals in such systems may be essentially corrupted by noises because of a great many independent noise sources and random variation of parameters of elements. Hence the problem of the best possible extraction of signals from noises arises. This problem can be solved only by applying optimal statistical information processing. Thus the statistical methods are necessary in certain cases for the shaping of control signals in automatic systems.

It is likely that such processes as learning, self-improvement and experience storage have a statistical nature. Thus statistical methods are also necessary for simulating processes of learning, self-improvement and experience storage in automatic systems.

Statistical methods yield also the solution of problems of the automatic check of control processes, in particular in industry, and problems of automatic detection and determination of defects with the purpose of their timely reparation.

It is therefore impossible to solve cardinal problems of automatic control without statistical methods, it is impossible to design many complex automatic systems without using statistical algorithms.

Yet the statistical methods are of a complex nature, and their application is generally bound with using high-speed computers. As a consequence, statistical methods are not yet so widely used in automatic systems design as more elementary deterministic methods. Statistical algorithms of information processing are often too complicated and cannot be used in automatic systems, owing to low speed and insufficient memory capacity of modern computers. As a consequence, statistical algorithms are seldom used in automatic systems for signal processing.

In spite of the difficulties connected with using statistical methods in the theory and the practice of automatic systems design, the field of application of statistical methods is continuously widening. As a result, the statistical theory of automatic control systems is now one of the most important sections in modern automatic control theory. With the development of analogue and digital computers the use of statistical methods becomes more general. With the increasing speed of computers statistical methods will be used more in automatic systems for signal processing.

Owing to a great variety of automatic control problems, which can be solved by statistical methods, the development of statistical methods in automatic control is proceeding in many directions. Thus it is scarcely possible to present a full survey of all directions of the development and the fields of application of statistical methods in automatic control in a brief report. Hence only the most important general statistical methods, according to the opinion of the author, will be considered here, many useful and necessary works in this field being omitted, for which the author hopes to be excused.

The main statistical problems in automatic control theory, which are being developed at present, are the following:

(1) The statistical analysis of systems with known dynamic characteristics.

(2) The statistical synthesis of automatic systems, their elements and optimal algorithms of control and information processing.

(3) The determining of dynamic characteristics (optimal models) of processes and plants.

(4) The statistical theory and methods of simulating learning processes, self-improvement and experience storage processes.

### Statistical Analysis of Linear Systems

The methods of the statistical theory of linear systems based on the correlation theory of random functions are the simplest statistical methods of automatic control systems analysis<sup>1</sup>. The system and its operator may be considered as deterministic, if all internal noises in its elements may be replaced by equivalent random inputs. To evaluate the accuracy of a linear system in such cases it is sufficient to know its matrix of weighting functions  $g(t, \tau)$ , the expectation (the mean value)  $m_x(t)$  and the correlation function  $K_x(t, t')$  of the vector of random inputs  $X(t)$ . Then the expectation  $m_y(t)$  and the correlation function  $K_y(t, t')$  of the vector of outputs of the system  $Y(t)$  are determined by

$$m_y(t) = \int_{t_0}^t g(t, \tau) m_x(\tau) d\tau \quad (1)$$

$$K_y(t, t') = \int_{t_0}^t \int_{t_0}^{t'} g(t, \tau) K_x(\tau, \sigma) \tilde{g}(t', \sigma) d\tau d\sigma \quad (2)$$

where  $\sim$  denotes a transposed matrix<sup>1, Sect. 90</sup>.

The calculations of eqns (1) and (2) may be generally accomplished with the aid of analogue computers after determining the weighting functions matrix  $g(t, \tau)$ . To determine weighting functions, the method of simulating adjoint systems may be used<sup>1-3</sup>.

Eqns (1) and (2) may be applied to continuous linear systems as well as to sampled-data ones. The weighting functions matrix  $g(t, \tau)$  is a linear combination of  $\delta$ -functions of  $\tau$  in the latter case, the coefficients being functions of  $t$  in general.

Another method for the evaluation of the accuracy of linear systems is based upon the expansion of random inputs in a series, whose members are uncorrelated elementary random functions (canonical expansion):

$$X(t) = m_x(t) + \sum_v V_v x_v(t) \quad (3)$$

where  $V_v$  are uncorrelated random variables (which are also statistically independent in the case of a normally distributed random function  $X$ ), and  $x_v(t)$  are certain definite functions, called coordinate functions. The canonical expansion of the form of eqn (3) may always be found for any random function (scalar or vector) in an infinite variety of ways, if its correlation function is given<sup>1, Ch. 9</sup>. In the case of the infinitesimal members of the expansion the sum in eqn (3) is replaced by the integral (integral canonical representation):

$$X(t) = m_x(t) + \int_{-\infty}^{\infty} V(\lambda) x(t, \lambda) d\lambda \quad (4)$$

where  $V(\lambda)$  is a random function of a parameter  $\lambda$  with the correlation function of the type of  $\delta$ -function (the generalized white noise), and  $x(t, \lambda)$ , considered as functions of  $t$  for various values of  $\lambda$ , are definite (non-random) coordinate functions.

The problem of shaping a given random function from a physical white noise using analogue computers represents the so-called problem of finding a shaping filter.

The generating of a given random function by passing a physical white noise through a shaping filter is thus a special way to find an integral canonical representation of this random function<sup>2</sup>. Vector random functions as well as scalar ones may be expressed by integral canonical representations, the coordinate functions  $x(t, \lambda)$ , as well as the coordinate functions  $x_v(t)$  of a canonical expansion (3), being vector functions in the previous case.

Expressing the random function  $X$  by the expansion (3) or the integral (4) and using the superposition principle, the evaluation of the accuracy of a linear system is reduced to the determination of its responses to the expectation of the input vector and to all coordinate functions  $x_v(t)$  or  $x(t, \lambda)$ . The correlation function of the output vector is then given by

$$K_y(t, t') = \sum_v D_v y_v(t) \tilde{y}_v(t') \quad (5)$$

or respectively by

$$K_y(t, t') = \int_{-\infty}^{\infty} G(\lambda) y(t, \lambda) \tilde{y}(t', \lambda) d\lambda \quad (6)$$

where  $D_v$  are variances of random variables  $V_v$ ,  $G(\lambda)$  is the intensity of the white noise  $V(\lambda)$  (i.e. the quantity by which the  $\delta$ -function is multiplied to give the correlation function of this white noise), and  $y_v(t)$ ,  $y(t, \lambda)$  are vector coordinate functions expressed as matrices-columns, representing the responses of the system to respective coordinate functions  $x_v(t)$ ,  $x(t, \lambda)$ , the bar denoting complex conjugates.

To determine the functions  $m_y(t)$ ,  $y_v(t)$ ,  $y(t, \lambda)$  in the case where the weighting functions matrix of the system is known, the formula (1) may be used, as well as analogous formulas

$$y_v(t) = \int_{t_0}^t g(t, \tau) x_v(\tau) d\tau \quad (7)$$

$$y(t, \lambda) = \int_{t_0}^t g(t, \tau) x(\tau, \lambda) d\tau \quad (8)$$

In the special case of a sampled-data system the weighting functions matrix  $g(t, \tau)$  in eqns (7) and (8) represent a linear combination of the  $\delta$ -functions of  $\tau$  whose coefficients are functions of  $t$  in the general case.

To determine the functions  $m_y(t)$ ,  $y_v(t)$ ,  $y(t, \lambda)$  in the case where the weighting functions matrix of the system is unknown, analogue computers or the linear system itself may be used. Passing in turn through a given system or its simulator the functions  $m_x(t)$ ,  $x_1(t)$ ,  $x_2(t)$ , ..., or the functions  $x(t, \lambda)$  corresponding to certain chosen values of  $\lambda$ , the respective functions  $m_y(t)$ ,  $y_1(t)$ ,  $y_2(t)$ , ..., or  $y(t, \lambda)$  are obtained.

It often occurs in practical problems, especially in those connected with automatic control in industry, that linear systems and their inputs are stationary or may be assumed to be stationary. The evaluation of systems accuracy is essentially simplified in this case, any stationary random function being representable by the integral canonical representation (4) with exponential coordinate functions (harmonic oscillations):

$$x(t, \lambda) = e^{i\lambda t} \quad (i = \sqrt{-1}) \quad (9)$$

The intensity  $G(\lambda)$  of the white noise  $V(\lambda)$  coincides in this case with the spectral density  $s_x(\lambda)$  of the stationary random function  $X$ <sup>1, Sect. 76</sup>. Similarly any stationary random vector function (i.e.

a vector random function with stationary and stationarily correlated components) may be expressed by the integral canonical representation in the form of the sum of the integrals of the form of eqn (4) with the vector coordinate functions

$$x(t, \lambda) = a(\lambda) e^{i\lambda t} \quad (10)$$

where  $a(\lambda)$  is a vector function of the parameter  $\lambda$ <sup>1, Sect. 78</sup>.

The coordinate functions  $y(t, \lambda)$  are easily determined in this case, because harmonic oscillations pass through a stable stationary linear system, being multiplied by the complex frequency response of the system. As a result the formula (6) shows that the output vector of the system is stationary, and its spectral and cross-spectral densities matrix is determined by the well-known formula

$$s_y(\omega) = \Phi(i\omega) s_x(\omega) \overline{\Phi(i\omega)} \quad (11)$$

where  $\Phi(\lambda)$  is the matrix of the frequency responses of the system<sup>1, Sect. 95</sup>. Formula (11) is evidently applicable only to asymptotically stable systems in a steady state. To determine variances and covariances of the outputs of a stationary linear system in a transient state or of an unstable system the general formula (6) must be used, in which the functions  $y(t, \lambda)$  are the transient responses of the system to the harmonic oscillations (9) or (10)<sup>1, Sect. 90, 92</sup>.

The theory of stationary random processes is thus insufficient in principle for basic automatic control problems, and only general theory of non-stationary random functions may serve as a mathematical basis of the statistical methods in automatic control theory.

The methods of evaluating the accuracy of linear systems with known dynamic characteristics may be considered at present as completely developed. The above formulae yield the precise solution of the problem for any linear system. Although these formulae require tedious calculations, these calculations are easily accomplished with the aid of analogue or digital computers.

Hence more convenient methods of calculations and more adequate algorithms now are the main subject of development in the theory of the statistical analysis of linear systems (see, for example, Nekolny and Benes<sup>4</sup>).

The knowledge of the first and second moments of system outputs is generally sufficient to characterize the accuracy of the system. In particular, the probability distribution of the outputs of a linear system is completely determined by these moments in the case of normally distributed inputs, because the distribution of outputs is also normal in this case.

If the input distribution is not normal, the knowledge of the first and second moments is then insufficient to determine the output distribution, and the moments of higher orders may be calculated using formulae analogous to (2), the respective moments of inputs must of course be known in this case.

It should be once more emphasized that the methods outlined above are applicable only to deterministic linear systems or to such systems whose internal noises and random variations of parameters can be replaced by equivalent random inputs. To determine the accuracy of linear systems with randomly changing dynamic characteristics, the statistical theory of systems reducible to linear should be used (see section on Precise Statistical Theory, and ref. 1, Sect. 106, 107 and ref. 5).

## The Approximate Methods of Statistical Analysis of Non-linear Systems

The statistical analysis of non-linear systems is an extremely difficult problem, and the precise theory of non-linear systems applicable to any non-linear systems does not exist at present.

The relative simplicity of the methods of statistical analysis of linear systems is the natural reason for trials to extend these methods for the problems of approximate analysis of accuracy of non-linear systems. Hence various methods of linearization of operators of non-linear systems have arisen.

The simplest form of linearization technique is based upon the expansion of all non-linear functions in the equations of a system (i.e. characteristics of all non-linear components) in Taylor series with respect to fluctuations, neglecting all members of the second and higher degrees. As a result, each non-linear function in the equations of the system  $\varphi(X)$  or  $\varphi(X_1, \dots, X_n)$  will be replaced by the approximate linear expression

$$\varphi(X) \approx \varphi(m_x) + \varphi'(m_x)(X - m_x) \quad (12)$$

or respectively

$$\begin{aligned} \varphi(X_1, \dots, X_n) &\approx \varphi(m_{x_1}, \dots, m_{x_n}) \\ &+ \sum_{v=1}^n \varphi'_{x_v}(m_{x_1}, \dots, m_{x_n})(X_v - m_{x_v}) \end{aligned} \quad (13)$$

where  $m_x, m_{x_1}, \dots, m_{x_n}$  are expectations of the random functions  $X, X_1, \dots, X_n$  respectively, just as in the previous section.

Using the formulae of the form of eqns (12) and (13), the equations of a system are linearized with respect to the fluctuations of the signals in various system components. This enables us to use the methods of the statistical theory of linear systems for approximate evaluation of the accuracy of non-linear systems. This gives non-linear equations for the expectations of signals, which coincide with the equations describing the behaviour of the system considered, as the expressions (12) and (13) are linear with respect to fluctuations only and are non-linear with respect to expectations<sup>1, Sect. 100-102</sup>.

The accuracy of the results obtained by the linearization technique based upon eqns (12) and (13) may be estimated by the magnitudes of errors of the expressions (12) and (13) in the range of practically possible values of the signal fluctuations.

Eqns (12) and (13) are valid only for continuous functions having continuous first derivatives with respect to all arguments in the domain of practically possible values of these arguments. Yet automatic systems often contain essentially non-linear components the characteristics of which are discontinuous (as, for instance, relays) or possess discontinuous derivatives (as, for instance, limiters). To linearize such characteristics the method of statistical linearization has been developed, proposed at first by Booton in U.S.A.<sup>6</sup> and by Kazakov in U.S.S.R.<sup>7, and 8, Ch. 3, 4</sup>.

The statistical linearization consists in replacing a non-linear component by a component linear with respect to fluctuations, conserving in a certain sense the magnitudes of output signal and output noise.

The most general form of the statistical linearization technique is based upon replacing all non-linear functions in the equations of a system by the approximate linear ones with respect to fluctuations expressions of the form<sup>9, and 8 Sect. 24, 25</sup>

$$\varphi(X(t)) \approx \varphi_0(t) + \int_{t_0}^t g(t, \tau) [X(\tau) - m_x(\tau)] d\tau \quad (14)$$

$$\text{or } \varphi(X_1(t), \dots, X_n(t)) \approx \varphi_0(t) + \sum_{v=1}^n \int_{t_0}^t g_v(t, \tau) [X_v(\tau) - m_{x_v}(\tau)] d\tau \quad (15)$$

The simplest form of the statistical linearization is Booton's statistical linearization, which is obtained from (14) by putting

$$\varphi_0(t) = k m_x(t), \quad g(t, \tau) = k \delta(t - \tau) \quad (16)$$

where  $k$  is determined so as to minimize the mean square error:

$$M[\{\varphi(X) - kX\}^2] = \min \quad (17)$$

$M[Z]$  denoting the expectation of the random variable  $Z$ .

Kazakov's statistical linearization in its first variant is obtained from (14) by putting

$$\varphi_0(t) = k_0 m_x(t), \quad g(t, \tau) = k_1 \delta(t - \tau) \quad (18)$$

and determining  $k_0$  and  $k_1$  so as to conserve the expectation and the variance of the output:

$$\left. \begin{aligned} M[k_0 m_x + k_1 (X - m_x)] &= M[\varphi(X)] \\ D[k_0 m_x + k_1 (X - m_x)] &= D[\varphi(X)] \end{aligned} \right\} \quad (19)$$

where  $D[Z]$  denotes the variance of the random variable  $Z$ . Kazakov proposed also to determine  $k_0$  and  $k_1$ , minimizing the mean square error:

$$M[\{\varphi(X) - k_0 m_x - k_1 (X - m_x)\}^2] = \min \quad (20)$$

Two conditions (19) and (20) giving a single value for  $k_0$  and two different values for  $k_1$ , Kazakov recommends the use of the arithmetic mean of these two values of  $k_1$ .

It should be emphasized that the previous statistical linearization of Booton and Kazakov based upon replacing a zero-memory non-linear component by a zero-memory linear one distorts the correlation function of the output and, as a consequence its spectrum in a stationary case. To conserve the spectrum of the output Pupkov proposed replacing a zero-memory non-linear component by a memory stationary linear system in a steady state, that is, to put  $g(t, \tau) = w(t - \tau)$ ,  $t_0 = -\infty$  in (14) and to determine  $w(\xi)$  (or corresponding transfer function) by equating the correlation functions of the left and right members of eqn (14)<sup>10</sup>.

An interesting generalization of Kazakov's statistical linearization is obtained by putting in (14)

$$g(t, \tau) = k_1 \delta(t - \tau) + l_1 \delta'(t - \tau) \quad (21)$$

The condition (19) is insufficient to determine  $k_1$  and  $l_1$  in this case. They may be determined by minimizing the mean square error

$$M[\{\varphi(X) - \varphi_0 - k_1 (X - m_x) - l_1 (\dot{X} - \dot{m}_x)\}^2] = \min \quad (22)$$

or by minimizing the mean square error under the additional condition of conserving the variance of the output:

$$D[\varphi_0 + k_1 (X - m_x) + l_1 (\dot{X} - \dot{m}_x)] = D[\varphi(X)] \quad (23)$$

The condition (22) yields the equations

$$\left. \begin{aligned} D_x k_1 + K_{xx} l_1 &= M[(X - m_x) \varphi(X)] \\ K_{xx} k_1 + D_x l_1 &= M[(\dot{X} - \dot{m}_x) \varphi(X)] \end{aligned} \right\} \quad (24)$$

for  $k_1$  and  $l_1$ , where  $D_x$ ,  $D_{xx}$  and  $K_{xx}$  are, respectively, variances of the random functions  $X$ ,  $\dot{X}$  and the covariance of their values corresponding with the same instant of time.

The conditional minimization of the mean square error conserving the variance of the output yields the same eqns (24) with the additional factor  $\lambda$  in their right members. This factor is determined by the condition (23) after solving eqns (24).

The third possible way to determine  $k_1$  and  $l_1$  is to determine them from the condition of conserving the variances of the output and its derivative.

All methods described above give for  $\varphi_0$  the formula

$$\varphi_0(t) = M[\varphi(X(t))] \quad (25)$$

In the case of the general statistical linearization according to eqn (14) Kazakov obtained from the condition of conserving the expectation and the correlation function of the output the formula (25) for the function  $\varphi_0(t)$  and the equation

$$\int_{t_0}^t \int_{t_0}^{t'} K_x(\tau, \sigma) g(t, \tau) g(t', \sigma) d\tau d\sigma = K_y(t, t') \quad (26)$$

for the weighting function  $g(t, \tau)$ , where

$$\begin{aligned} K_y(t, t') &= M[\varphi(X(t)) \varphi(X(t'))] \\ &\quad - M[\varphi(X(t))] M[\varphi(X(t'))] \end{aligned} \quad (27)$$

is the correlation function of the output of the non-linear component<sup>8, Sect. 24, 25, and 9</sup>. The minimization of the mean square error yields in this case the same formula (25) for  $\varphi_0$  and the equation

$$\int_{t_0}^t g(t, \tau) K_x(\tau, \sigma) d\tau = K_{yx}(t, \sigma) \quad (t_0 \leq \sigma \leq t) \quad (28)$$

for the weighting function  $g(t, \tau)$ , where

$$K_{yx}(t, \sigma) = M[\{X(\sigma) - m_x(\sigma)\} \varphi(X(t))] \quad (29)$$

is the cross-correlation function of the output and input of the non-linear component<sup>8, Sect. 24, 25, and 9</sup>.

It should be noted, that in the case of a single-valued function  $\varphi$  Booton's and Kazakov's first statistical linearization, minimizing the mean square error gives as a special case the harmonic linearization of a non-linear component, provided that the sinusoidal input is assumed as random with a definite amplitude and uniformly distributed over the period random phase. This is easily seen from the works of Pervozvansky<sup>11</sup>.

More general statistical linearization according to formulas (14) and (21) minimizing the mean square error gives as a special case the harmonic linearization of a non-linear component in the case of a multi-valued characteristic  $\varphi(X)$  as well (hysteresis component, for instance).

The above formulae show, that the statistical gains of a non-linear component  $k$ ,  $k_0$ ,  $k_1$ ,  $l_1$  as well as its statistical weighting function  $g(t, \tau)$  depend on the probability distribution of input, which cannot be known *a priori*, before the evaluation of the system accuracy. To determine these quantities Booton and Kazakov recommend assuming the normal distribution for inputs of non-linear components. Then the coefficients  $k$ ,  $k_0$ ,  $k_1$ ,  $l_1$  will be functions of the unknown expectations and variances of inputs, and the application of the statistical linearization and



the methods of the statistical theory of linear systems yields a sufficient set of equations to determine all quantities that are unknown<sup>1, Sect. 103-105; 8, Sect. 32, 34-36; 12, Sect. 11.2, 11.3</sup>.

The statistical linearization of KAZAKOV has been generalized by SOMMERVILLE and ATHERTON for the case where the input of a non-linear component is the sum of independent random signals with different probability distributions<sup>13</sup>. Their statistical linearization in a generalized form is expressed by the formulae

$$X(t) = \sum_{v=1}^n X_v(t) \quad (30)$$

$$\varphi(X(t)) \approx \varphi_0(t) + \sum_{v=1}^n \int_{t_0}^t g_v(t, \tau) [X_v(\tau) - m_{x_v}(\tau)] d\tau \quad (31)$$

Sommerville and Atherton have considered the case, which is obtained from (31) by putting

$$\varphi_0(t) = k_0 m_x(t), g_v(t, \tau) = k_v \delta(t - \tau) \quad (v = 1, \dots, n) \quad (32)$$

The condition of conserving the variance of the output is evidently insufficient for the determination of the weighting functions  $g_v(t, \tau)$  or, in Sommerville and Atherton's case, of the coefficients  $k_v$ . So the minimization of the mean square error is expedient for this purpose. Using this criterion, Sommerville and Atherton obtained the expressions for the coefficients  $k_v$  in (32). It should be noted, that the coefficients  $k_v$  may be determined in this case also by the conditional minimizing of the mean square error conserving the variance of the output.

An interesting generalization of Sommerville and Atherton's statistical linearization is obtained from (33) by putting

$$g_v(t, \tau) = k_v \delta(t - \tau) + l_v \delta'(t - \tau) \quad (v = 1, \dots, n) \quad (33)$$

Then the statistical linearization gives as a special case the joint statistical and harmonic linearization, if the input of a non-linear component is the sum of a sinusoidal signal with a definite amplitude and uniformly distributed over the period random phase and a normally distributed noise. This joint statistical and harmonic linearization is generally used to determine the effect of noises upon the parameters of oscillations of non-linear systems<sup>12, Sect. 10.7; 14, Ch. 10; 8, Sect. 38; 11, Sect. 3.2, 3.3</sup>.

All the methods outlined above yield the same expression (25) for  $\varphi_0$ .

For the weighting functions  $g_v(t, \tau)$  in (31) the minimizing of the mean square error yields the system of linear integral equations, which may be written using matrix notations in the form of eqn (28).

The development of the statistical linearization method and its applications have been given in numerous papers by scientists of various countries. Among these the work of Japanese scientists Sawaragi, Sunahara and others should be especially noted<sup>15-19</sup>.

Based upon various forms of linearization, approximate methods of the analysis of non-linear systems accuracy are the simplest and the most effective methods of the statistical theory of non-linear systems, owing to which they have been widely expanded. It is impossible to give even the enumeration of all contributions to this field in this brief report.

The natural way to increase the accuracy of the statistical methods of non-linear systems analysis based upon linearization is to use the members of the second and higher degrees in the series expansions of non-linear functions. So the method of

series expansions of outputs of non-linear systems with respect to random parameters<sup>1, Sect. 108</sup> and DOSTUPOV's method of equivalent disturbances<sup>8, Ch. 11, and 20</sup> have arisen. The latter is based upon simulating some typical realizations of random inputs to a system chosen in such a way, that some linear combinations of various degrees of resulting system outputs with definite coefficients be equal to corresponding moments of system outputs. Owing to this Dostupov's method is especially fit for the evaluation of non-linear systems accuracy with the aid of digital computers.

Another method of non-linear systems statistical analysis based upon the equation in partial derivatives determining the first probability distribution of system outputs has been developed<sup>21</sup>.

Besides the general approximate methods of statistical non-linear systems analysis outlined above, other methods have been applied to solve some special problems. We shall confine ourselves to mentioning the papers by Feldbaum<sup>22, 23</sup> and by Perovzovskiy<sup>24, 25, also 11, Ch. 4</sup>), who have applied the theory of Markov chains to study the process of extremum seeking in sampled-data systems of optimized control, and papers by Krasovsky, who has obtained the approximate differential equations for the expectations of the coordinates of continuous systems with extremal parameters adjustment<sup>26-28</sup>.

### The Precise Statistical Theory of Systems Reducible to Linear

The statistical theory of linear systems is also applicable to certain non-linear systems which form a class of systems reducible to linear<sup>1, Sect. 106, 107</sup>. This class contains, in particular, all systems obtained by cascading and joining in parallel linear systems, multipliers, devices for raising in various degrees and certain types of functional devices, provided that non-linear components are not contained in feedback loops.

A system is called reducible to linear (in the statistical sense) if its input-output relation is of the form

$$Y(t) = \int_{t_0}^t \int_{t_0}^t g(t, \tau_1, \dots, \tau_k) \varphi(X(\tau_1), \dots, X(\tau_k), \tau_1, \dots, \tau_k, t) d\tau_1, \dots, d\tau_k \quad (34)$$

where  $g(t, \tau_1, \dots, \tau_k)$  is a weighting functions matrix, and  $\varphi(X(\tau_1), \dots, X(\tau_k), \tau_1, \dots, \tau_k, t)$  is a definite vector function of the values of input vector  $X(t)$  corresponding to time instants  $\tau_1, \dots, \tau_k$ , which may also depend on  $\tau_1, \dots, \tau_k, t$ .

Putting

$$U(t, \tau_1, \dots, \tau_k) = \varphi(X(\tau_1), \dots, X(\tau_k), \tau_1, \dots, \tau_k, t) \quad (35)$$

the output vector of the system  $Y(t)$  is expressed as a result of a linear transform of the random vector function  $U$ :

$$Y(t) = \int_{t_0}^t \dots \int_{t_0}^t g(t, \tau_1, \dots, \tau_k) U(t, \tau_1, \dots, \tau_k) d\tau_1, \dots, d\tau_k \quad (36)$$

Knowing 2  $k$ -dimensional probability density of the input vector (i.e. the joint probability density of values of all components of the vector  $X(t)$ , corresponding to arbitrary time instants  $\tau_1, \dots, \tau_k, \tau'_1, \dots, \tau'_k$ ), the expectation and the correlation function of the vector random function  $U$  is defined by the well-known formulae of the probability theory, after which the expectation and the correlation function of the output vector of

the system  $Y(t)$  may be determined by the methods considered in the first section.

Any system whose output vector is a functional polynomial (i.e. a truncated functional Taylor series) with respect to components of its input vector is evidently reducible to linear<sup>29</sup>.

Theory of certain special classes of systems reducible to linear has been developed by Zadeh<sup>30, 31</sup>, Lubbock<sup>32, 33</sup> and Prasad<sup>34</sup>. The general theory of systems reducible to linear has been given elsewhere<sup>1, 5</sup>.

As any system may be replaced (theoretically at least) by a sufficiently near system reducible to linear<sup>30, 31</sup>, the statistical theory of systems reducible to linear may be used for approximate evaluation of the accuracy of any non-linear systems.

To apply the theory of systems reducible to linear to statistical analysis of linear systems with randomly varying dynamic characteristics it is sufficient to consider the matrix of weighting functions of a linear system  $G(t, \tau)$  as a random function<sup>1, Sect. 98</sup>.

Another somewhat less general approach to the statistical theory of non-linear systems has been given by Wiener<sup>35</sup>.

### The Precise Statistical Theory of One Class of Non-linear Systems Based upon the Theory of Markov Processes

Besides considering in the previous section systems reducible to linear, the precise statistical analysis is now available only for such systems, which are described by ordinary differential equations and inputs of which are either white noises or are expressed in terms of white noises by means of ordinary differential equations of finite order (random disturbances, which are stationary random functions with rational spectral densities are an example).

More precisely, the class of non-linear systems, which can be treated by the precise statistical theory based upon the theory of Markov random processes, is defined as the class of all systems described by systems of ordinary differential equations of the finite order of the form

$$\dot{X} = \varphi(t, X) + \psi(t, X) V(t) \quad (37)$$

where  $X$  is the vector of coordinates of the system, the components of which are all the outputs of the system, all the necessary auxiliary variables and those of the system inputs which are not white noises,  $V(t)$  is the vector random function whose components are white noises in a strict sense\*,  $\varphi(t, X)$  is a vector function of the time  $t$  and the vector  $X$ ,  $\psi(t, X)$  is a matrix whose elements may be functions of  $t$  and  $X$ .

The vector of system coordinates  $X$  defined by the system of differential equations (37), subject to conditions indicated, represents a Markov random process, and its first probability distribution is determined by the well-known equations of the theory of Markov processes.

The theory of Markov random processes was used at first for analysing dynamic systems accuracy by Andronov, Vitt and Pontryagin<sup>36</sup>. Further development of the statistical theory of non-linear systems, based upon the theory of Markov processes, has been given by Barrett<sup>37</sup>, Pugachev<sup>38, 39</sup>, Hazen<sup>40, 41</sup> and others.

As shown<sup>38, 39</sup>, the first characteristic function  $g_1(\lambda; t)$  of

the vector of system coordinates  $X(t)$  defined by (37) is determined by the linear integro-differential equation

$$\frac{\partial g_1(\lambda; t)}{\partial t} = \frac{1}{(2\pi)^n} \int_{-\infty}^{\infty} dx \int_{-\infty}^{\infty} e^{i(\lambda - \mu)x} \Phi(\lambda; t, x) g_1(\mu; t) d\mu \quad (38)$$

where  $n$  is the number of components of the vector  $X(t)$ ,  $x$ ,  $\lambda$  and  $\mu$  are  $n$ -dimensional vector arguments,  $(\lambda - \mu)x$  is the scalar product of the vectors  $\lambda - \mu$  and  $x$ , and  $\Phi(\lambda; t, x)$  is the value at  $\Delta t = +0$  of the derivative with respect to  $\Delta t$  of the conditional characteristic function of the increment of the process  $X(t)$  during the time  $\Delta t$ , when its value  $x$  at  $t$  is given:

$$\Phi(\lambda; t, x) = i\lambda \varphi(t, x) + \kappa(\lambda; t, x) \quad (39)$$

$$\kappa(\lambda; t, x) = \lim_{\Delta t \rightarrow +0} \frac{1}{\Delta t} \left\{ M \left[ \exp \left\{ i\lambda \cdot \int_t^{t+\Delta t} \psi(\tau, x) V(\tau) d\tau \right\} \right] - 1 \right\} \quad (40)$$

(the point denotes the scalar product).

Under the same conditions the first probability density  $f_1(x; t)$  of the process  $X(t)$  is determined by the linear integro-differential equation<sup>39</sup>

$$\frac{\partial f_1(x; t)}{\partial t} = \frac{1}{(2\pi)^n} \int_{-\infty}^{\infty} d\lambda \int_{-\infty}^{\infty} e^{i\lambda(\xi - x)} \Phi(\lambda; t, \xi) f_1(\xi; t) d\xi \quad (41)$$

In the special case, in which the first and the second moments of the random vector

$$Z = \int_t^{t+\Delta t} V(\tau) d\tau \quad (42)$$

are of order  $\Delta t$ , and all higher moments are infinitesimal of higher orders, the eqn (41) is reduced to the well-known Fokker-Plank-Kolmogorov equation.

In another special case, where the function  $\Phi(\lambda; t, x)$  is a polynomial with respect to components of the vector  $x$ , eqn (38), defining the characteristic function of the system coordinates, is reduced to the linear equation in partial derivatives

$$\frac{\partial g_1(\lambda; t)}{\partial t} = \Phi \left( \lambda; t, \frac{\partial}{\partial \lambda} \right) g_1(\lambda; t), \quad (43)$$

where  $\partial/\partial \lambda$  is the vector gradient in the space of the vector  $\lambda$ .

In the special case of a linear system, eqn (43) is of the first order and is easily integrated by the well-known standard procedure.

The solving of eqns (38) and (41) for  $g_1(\lambda; t)$  and  $f_1(x; t)$  (in respective special cases eqn (43) or Fokker-Plank-Kolmogorov equation) may be accomplished analytically only in the simplest problems. Thus, only numerical solution by means of digital computers is available in a great majority of problems. But the calculations required are so tedious, that it is possible to accomplish them with the aid of electronic digital computers, only when the order  $n$  of the system of differential equations (37) (i.e. the number of degrees of freedom of the system) isn't too high. The volume of calculations involved and the necessary storage capacity of computers increase rapidly with the increasing of the order of differential equations. That is the reason why the theory of Markov random processes doesn't yet find sufficiently wide applications in automatic control problems.

\* White noises in a strict sense are such random functions whose values for different values of the argument are independent and whose correlation and cross-correlation functions have as a factor  $\delta$ -function of the difference of arguments.

### The Problem of Finding an Adequate Dynamic Model of a Controlled Process or a Plant

One of the most important problems of the statistical theory of automatic control systems is the determining of dynamic characteristics of systems and plants from normal operating records or, more precisely, finding adequate models of systems and plants. To solve this problem it is necessary to find, by a suitable statistical processing of normal operating records, the correlation function of the input vector to the system under study  $K_x(t, t')$  and the cross-correlation function of the output and input vectors  $K_{yx}(t, t')$ . Then the weighting functions matrix of the linear system or plant is determined by solving the set of linear integral equations of the form of eqn (28)<sup>42, 43</sup>.

This method enables us to determine dynamic characteristics of linear systems subject to noise effects with better accuracy, than the usual methods of determining responses to standard types of inputs.

The eqn (28) also defines the weighting functions matrix of the linear system yielding the best approach, in the mean square error sense, to a given non-linear system (i.e. the best linear model of a given non-linear system).

### Methods of Determining Optimal Linear Systems

The problem of determining characteristics of systems from normal operating records is a special case of more general problems of the statistical theory of optimal systems. This theory enables us to find optimal mathematical operations over inputs providing the best attainable accuracy for given statistics of signals and noises. Thus, the theory of optimal systems affords finding the theoretical limit of the accuracy of a system of a given destination. Comparing with this theoretical limit the accuracy of the system to be designed, the engineer can estimate the accuracy of this system and provide the nearness of the system accuracy to the optimal one.

The general problem of the statistical theory of optimal systems may be stated in the following way: it is necessary to find the system operator  $A$  in such a manner, that for a given input  $Z(t)$ , representing a signal distorted by noise, the system output  $W^*(t) = AZ(t)$  be as near as possible to a desired output  $W(t)$ , the statistics of the input and desired output being given. This problem covers all problems of the determining of optimal systems and optimal information processing algorithms, arising in automation, as well as a variety of analogous problems, arising in other branches of science and engineering. Such are, for example, the problems of optimal detection of various signals in noises, arising in radio-engineering, seismology and other branches of engineering; the problem of the best reproduction of the sound and TV picture; the determining and the extrapolation of the elements of trajectories of satellites and cosmic rockets from the observed data, etc.

To estimate the nearness of the system output  $W^*$  to the desired output  $W$ , various criteria are used depending on the destination of the system. Almost all practically employed criteria admit representation in the form of a condition of minimum average risk:

$$\varrho = M [I(W, W^*)] = \min \quad (44)$$

where  $I$  is a certain function (or functional) of desired and actual systems outputs, generally called the loss function. The criteria

of the form (44) are generally called Bayesian criteria. The optimal algorithm of the system, processing output  $W^*$  for a given input  $Z$  in accordance with the criterion (44), is called Bayesian decision.

For automatic control systems the criterion of the minimum of the mean square error is most widely used, to which corresponds the quadratic loss-function

$$I(W, W^*) = (W^* - W)^2 \quad (45)$$

Although this criterion is the simplest from a mathematical viewpoint, it isn't adequate in many problems. Other criteria must then be used. So, for example, in the problem of signal detection in noise the criterion of the minimum probability of the error is expedient, sometimes with certain additional conditions. This criterion admits the representation in the form of eqn (44), where the loss function  $I$  is zero for the correct decision and has definite positive values for errors of two kinds (the decision that the signal is absent, when it is present, and the decision that the signal is present, when only the noise enters the system). In more complex situations the loss-function may depend not only on the present values of the desired and actual outputs, but may be a functional of  $W$  and  $W^*$ .

In many practical problems a signal contained in the system input  $Z$  has a regular component, which is a known function of time and the finite number of parameters  $U_1, \dots, U_N$ . These parameters are random variables or unknown non-random quantities which may have any value. The desired output is then the result of a given transform of the signal contained in the input  $Z$ .

If the regular part of the signal is a linear function of parameters  $U_1, \dots, U_N$ , and the desired output represents a result of a given linear transform of the signal, then

$$Z(t) = \sum_{r=1}^N U_r \varphi_r(t) + X(t) \quad (46)$$

$$W(t) = \sum_{r=1}^N U_r \psi_r(t) + Y(t) \quad (47)$$

where  $\varphi_r(t)$ ,  $\psi_r(t)$  ( $r = 1, \dots, N$ ) are known functions of time,  $X(t)$  is the sum of the irregular part of the signal and the noise, and  $Y(t)$  is the result of a given linear transform of the irregular part of the signal. The problem of finding the optimal linear system, using the minimum mean square error criterion, is then reduced to solving  $N + 1$  systems of linear integral equations

$$\int_{t-T}^t g^{(0)}(t, \tau) K_x(\tau, \sigma) d\tau = K_{yx}(t, \sigma) \quad (t - T \leq \sigma \leq t) \quad (48)$$

$$\int_{t-T}^t g^{(r)}(t, \tau) K_x(\tau, \sigma) d\tau = \varphi_r(\sigma) \quad (49)$$

$$(t - T \leq \sigma \leq t; r = 1, \dots, N)$$

and the system of linear algebraic equations

$$\sum_{q=1}^N (C_{pq} + c_{pq}) \lambda_q = \psi_p(t) - C_{p0} \quad (p = 1, \dots, N) \quad (50)$$

where  $C_{pq}$  and  $C_{p0}$  are defined by

$$C_{pq} = \int_{t-T}^t g^{(q)}(t, \tau) \varphi_p(\tau) d\tau = \int_{t-T}^t g^{(p)}(t, \tau) \varphi_q(\tau) d\tau \quad (51)$$

$$(p, q = 1, \dots, N)$$

$$C_{p0} = \int_{t-T}^t g^{(0)}(t, \tau) \varphi_p(\tau) d\tau = \int_{t-T}^t K_{yx}(t, \tau) g^{(p)}(t, \tau) d\tau$$

$$(p = 1, \dots, N) \quad (52)$$

and  $C_{pq}$  are the elements of the matrix, inverse with respect to the matrix of second moments of the random variables  $U_1, \dots, U_N$ :

$$\gamma_{pq} = M[U_p U_q] \quad (p, q = 1, \dots, N) \quad (53)$$

The product of vector functions  $g^{(q)}(t, \tau)$  and  $\varphi_p(\tau)$  in (51) is the scalar product.

If  $U_r$  is a non-random unknown quantity which may have any value, then all  $C_{pq}$ , for which either  $p$  or  $q$  is equal to  $r$ , are equal to zero.

After finding the weighting functions matrix  $g^{(0)}$ , the vector functions  $g^{(1)}, \dots, g^{(N)}$  and the vectors  $\lambda_1, \dots, \lambda_N$ , the weighting functions of the optimal system are given by

$$g_{hi}(t, \tau) = g_{hi}^{(0)}(t, \tau) + \sum_{r=1}^N \lambda_{rh} g_i^{(r)}(t, \tau) \quad (54)$$

The mean square of the error of the  $h$ th output of the optimal system is given by

$$\eta_{h \min} = D_{y_h}(t) - \sum_{t-T}^t g_{hi}^{(0)}(t, \tau) K_{y_h x_i}(t, \tau) d\tau$$

$$+ \sum_{r=1}^N \lambda_{rh} [\psi_{rh}(t) - (C_{r0})_h] \quad (55)$$

In the special case, where the signal has no irregular component,  $Y(t) \equiv 0$ ,  $D_{y_h}(t) \equiv 0$ ,  $K_{yx}(t, \tau) \equiv 0$ ,  $g^{(0)}(t, \tau) \equiv 0$ ,  $C_{10} = \dots = C_{N0} = 0$ , and eqns (54) and (55) are accordingly simplified<sup>1, Sect. 120, 122, 124, 125</sup>.

Andreev has shown that the problem of finding the optimal system, using a more general criterion of the extremum of any pre-assigned function of the expectation and the variance of the system error, is reduced to equations of the same type, as above<sup>44-46</sup>.

Thus, the problem of determining the optimal linear system, using the mean square error criterion or a more general criterion of extremum of a given function of the expectation and the variance of the system error, is reduced to solving the system of linear integral equations of the form

$$\int_{t-T}^t g(t, \tau) K_x(\tau, \sigma) d\tau = f(t, \sigma) \quad (t-T \leq \sigma \leq t) \quad (56)$$

where  $f(t, \sigma)$  is a known matrix, and  $g(t, \tau)$  is the unknown matrix of weighting functions. These equations define the weighting functions of a continuous linear system in the general case. Thus, the optimal system among all linear systems is a continuous linear system in general. To find the optimal sampled-data linear system, it is necessary to search its weighting functions in the form of linear combinations of  $\delta$ -functions. Eqn (56) is then reduced to a set of linear algebraic equations.

Eqn (28), encountered in the statistical linearization problem as well as in the problem of estimating the dynamic characteristics of systems from normal operating records, is evidently a special case of eqn (56).

All known methods of solving the equations of the form of eqn (56) are in fact based upon the method of canonical representations of random functions. The method of integral

canonical representations yields the solution of such equations in a closed form in all cases, where an integral canonical representation of the random function  $X$  can be found<sup>1, Sect. 128-133</sup>. The method of canonical expansions provides an approximate solution in the form of truncated series in all cases, as any random function can always be represented by a canonical expansion in an infinite number of ways<sup>1, Sect. 134-137</sup>.

To solve eqn (56) by the integral canonical representations method, when the observation interval  $T$  is infinite, it is necessary to find a linear system transforming the vector random function  $X$  to a corresponding number of uncorrelated white noises. The weighting functions matrix  $w^-(t, \tau)$  of such a system have been found, the solution of eqn (56) for  $T = \infty$  is given by

$$g(t, \tau) = \int_{-\infty}^t d\lambda \int_{-\infty}^{\lambda} f(t, \sigma) \tilde{w}^-(\lambda, \sigma) G^{-1}(\lambda) w^-(\lambda, \tau) d\sigma \quad (57)$$

where  $G(\lambda)$  is the diagonal matrix of intensities of the components of the vector white noise, to which the vector random function  $X$  is transformed by the system with the weighting functions matrix  $w^-(t, \tau)$ .

In the special case, where the correlation function  $K_x(\tau, \sigma)$  depends only on the difference  $\tau - \sigma$  (i.e. the vector random function  $X$  is stationary), and the right member of eqn (56)  $f(t, \sigma)$  is dependent only on the difference  $t - \sigma$  [for instance, a cross-correlation function of two stationary and stationarily correlated random functions  $Y$  (a desired stationary vector of system outputs) and  $X$  (a stationary input vector)],  $f(t, \sigma) = h(t - \sigma)$ , eqn (57) yields the known formula of Wiener's theory of smoothing, interpolation and extrapolation of stationary random processes<sup>1, Sect. 129, and 48, 49</sup>:

$$\Psi(\lambda) = \frac{1}{2\pi} \int_0^{\infty} d\xi \int_{-\infty}^{\infty} \psi(\mu) \tilde{\Phi}^{-1}(i\mu) \Phi^{-1}(\lambda) e^{(i\mu - \lambda)\xi} d\mu \quad (58)$$

where  $\Psi(\lambda)$  is the matrix of transfer functions of the optimal stationary linear system,  $\psi(\mu)$  is given by

$$\psi(\mu) = \frac{1}{2\pi} \int_{-\infty}^{\infty} h(\eta) e^{-i\mu\eta} d\eta \quad (59)$$

and  $\Phi^{-1}(\lambda)$  is the square matrix of transfer functions of the stationary stable minimal-phase system transforming the vector random function  $X$  to the vector white noise with uncorrelated components having unit spectral densities (i.e. the intensities equal to  $2\pi$ ). The special case of Wiener's formula for a single-input and single-output system has been generalized by Booton for an arbitrary function  $f(t, \sigma)$ <sup>50</sup>.

The general formula (57) may also be used to find the solution of eqn (56) in the case of an arbitrary finite observation interval  $T$ . For this purpose it is sufficient to let  $g(t, \tau) = 0$  for  $\tau < t - T$  in (57), and to assume  $f(t, \tau)$  as unknown in the interval  $-\infty < \tau < t - T$ . This yields simultaneous equations determining the matrix  $g(t, \tau)$  in the interval  $t - T \leq \tau \leq t$  and the matrix  $f(t, \tau)$  in the interval  $-\infty < \tau < t - T$ . These equations are easily solved in the case, where the vector random function  $X$  is stationary and has a rational matrix of spectral and cross-spectral densities (Zadeh and Ragazzini's case<sup>51</sup> and its generalization<sup>48</sup>), and in a more general case, where the vector random function  $X$  can be expressed in terms of a vector white noise with a respective number of uncorrelated components by simultaneous linear differential equations with sufficiently smooth coefficients<sup>1, Sect. 130-133; 48, and 2, Sect. 8-5</sup>.

To solve eqn (56) by the canonical expansions method it is necessary to find a canonical expansion of the vector random function  $X(\tau)$  of the form of eqn (3) in the observation interval  $t - T \leq \tau \leq t$ . Then vector functions  $a_v(\tau)$  will be found, satisfying, together with the vector coordinate functions  $x_v(\tau)$ , the equations

$$\int_{t-T}^t x_v(\tau) \overline{a_\mu(\tau)} d\tau = \delta_{v\mu}$$

$$x_v(\tau) = \frac{1}{D_v} \int_{t-T}^t K_x(\tau, \sigma) a_v(\sigma) d\sigma \quad (60)$$

Then the solution of eqn (56) is defined by<sup>1</sup>, Sect. 134-137

$$g(t, \tau) = \sum_v \overline{\frac{a_v(\tau)}{D_v}} \cdot \int_{t-T}^t f(t, \sigma) a_v(\sigma) d\sigma \quad (61)$$

where the point means the product of a vector column by a vector row. Eqn (61) represents the solution of eqn (56) in all cases, where the right member of eqn (56) can be expressed in the observation interval  $t - T \leq \tau \leq t$  by the coordinate-functions expansion

$$f(t, \tau) = \sum_v \overline{x_v(\tau)} \cdot \int_{t-T}^t f(t, \sigma) a_v(\sigma) d\sigma$$

where the point has the same meaning as in eqn (61). This condition (which is equivalent to the condition of the completeness of the set of coordinate functions  $x_v(\tau)$  in the observation interval  $t - T \leq \tau \leq t$ ) is generally fulfilled in practical problems.

Eqn (61), in particular, may serve for determining optimal sampled-data linear systems. The random function  $X$  should be expressed by the expansion (3) for all the instants of application of inputs to the system in this case. The functions  $a_v(\tau)$  being linear combinations of  $\delta$ -functions in this case, the function  $g(t, \tau)$  determined by eqn (61) will also be a linear combination of  $\delta$ -functions, i.e. it will represent the weighting functions matrix of a sampled-data linear system. For a finite number of instants of the application of inputs to a system, eqn (61) yields the precise solution of corresponding simultaneous linear algebraic equations determining the optimal sampled-data linear system.

The above outlined methods of finding optimal linear systems cover practically all problems of this kind encountered in automation, radio engineering, and in other branches of sciences and engineering. In its general form the canonical representations method is also applicable to problems in which the observed signal (input vector)  $Z$ , as well as the desired and actual outputs  $W$ ,  $W^*$  are random functions of several variables. Eqn (61), in particular, is also valid in this case, the variables  $t$ ,  $\tau$ ,  $\sigma$  being vectors, and the integrals being multiple and extended over the corresponding domain of the space of the vectors  $\tau$  and  $\sigma$  (the observation domain).

The common deficiency of all methods outlined above is that they determine only overall system weighting functions and provide no solution of optimal systems synthesis problem, when some invariable components of the system (plants for instance) are given. Owing to this, various attempts have been made to find other methods allowing a direct optimal linear systems synthesis.

The most interesting and effective results in this direction have been obtained by Kalman and Bucy<sup>52, 53</sup>. Their approach yields directly differential equations of the optimal system. Un-

fortunately this method is restricted to problems, where the interferences are white noises, and the signal is formed from white noises passing through a certain linear system during the observation interval. Owing to this, Kalman and Bucy's method can be used only for the synthesis of linear systems with infinite memory and with white noises as interferences.

Another interesting approach to the problem of the direct optimal linear system synthesis is formed by attempts to extend the theory of analytical design of controllers to systems with random inputs<sup>54</sup>.

The statistical problem of finding optimal control was at first stated and solved under certain assumptions by Pontryagin<sup>55</sup>. This theory has been developed further by Krasovskiy<sup>56, 57</sup>. Yet the common deficiency of all papers, in which the theory of analytical design of controllers is extended to systems with random inputs, is the assumption that random disturbances are directly measurable, whereas neither signals nor noises can be measured separately in general. This fact being taken into account in Zaytsev's papers<sup>58</sup>, a certain generalization of Wiener's problem of finding optimal stationary linear system has been obtained as a result.

The statistical theory of optimal linear systems is also applicable to systems reducible to linear. An effective solution of the problem of finding the optimal non-linear system among systems reducible to linear was at first given in 1956<sup>59</sup>. General methods of solution of this problem, stated in all possible ways, have also been given in<sup>1</sup>, Sect. 138, 139, and 5. The equation determining the optimal system for one of the possible statements of the problem was previously obtained by Zadeh<sup>30, 31</sup>. The solutions of some concrete problems of finding optimal systems reducible to linear have been given by Lubbock<sup>32, 33</sup> and Prasad<sup>34</sup>.

### The Application of Statistical Decision Theory to Problems of Finding Optimal Systems and Algorithms of Information Processing

As stated in the Introduction, the optimal statistical information processing is of great importance in the design of complex systems containing many elements. The general problem of the statistical theory of optimal systems, stated in the beginning of the preceding section, covers all problems of finding optimal algorithms of automatic systems taken overall and of their various components as well. Hence the methods of solving this problem may serve for estimating the potential accuracy and potential interference-stability of overall automatic systems and for designing optimal components of automatic systems as well. In particular, they may be used in the design of systems for the detection and extraction of signals in noises.

The general problem of the theory of optimal systems, stated in the beginning of the preceding section, is a typical mathematical problem of the statistical decision theory. Hence the natural theoretical basis for the solution of problems of this kind is given by the statistical decision theory initiated by Wald<sup>60</sup> and primarily used for solving problems of detection and extraction of signals in noises by Middleton and van Meter<sup>61-63</sup>. The effective general method of solving a wide class of problems of the statistical decision theory, applicable to the majority of problems arising in automatic control theory, has been<sup>1</sup>, Sect. 141-145, and 64-70 developed.

Let  $S$  be an input signal to the system or its component to be designed. This signal is corrupted by noises, owing to which

the input  $Z$  enters the system instead of  $S$ . The ensemble of possible realizations of the signal  $S$  is characterized by *a priori* probability density  $f(s)$ .

The effect of noises is that the input  $Z$  is random for each realization  $s$  of the signal  $S$ . Hence the effect of noises is characterized by the conditional probability density  $f_1(z|s)$  of the input  $Z$  with respect to the signal  $S$ .

The desired output of the system (or its component) to be designed  $W$  is usually the result of a certain pre-assigned transform of the signal  $S$ . The statistical decision theory admits, in general, that this transform may be random and consequently may be defined by the conditional probability density  $f_2(w|s)$  of the desired output  $W$  with respect to the signal  $S$ .

The optimal algorithm of the system (or its component) to be designed may also be random in the general case. So it will be completely determined, if the conditional probability density  $\delta(w^*|z)$  of the output  $W^*$  with respect to the input  $Z$  is found. The probability density  $\delta(w^*|z)$  or the corresponding probability measure is called the *statistical decision function*.

We emphasize, that the desired output  $W$  is dependent on the signal  $S$  only and is independent of noises. Hence  $f_2(w|s)$  is identically equal to the conditional probability density of  $W$  with respect to  $S$  and  $Z$ :  $f_2(w|s) \equiv f_3(w|s, z)$ . Similarly the actual output  $W^*$  is dependent only on the input  $Z$  and cannot depend on  $S$  and  $W$ . Hence the statistical decision function  $\delta(w^*|z)$  is identically equal to the conditional probability density of  $W^*$  with respect to  $S$ ,  $Z$  and  $W$ :  $\delta(w^*|z) \equiv \delta_1(w^*|s, z, w)$ .

Using these facts, the joint probability density of  $S$ ,  $Z$ ,  $W$  and  $W^*$  is given by

$$f_0(s, z, w, w^*) = f(s)f_1(z|s)f_2(w|s)\delta(w^*|z) \quad (62)$$

where probability densities  $f$ ,  $f_1$ ,  $f_2$  are known, and  $\delta$  is to be determined from the condition of the minimum of the average risk (44), which is evidently dependent on  $\delta$ . Putting this in evidence and using (62), we obtain the following expression for the average risk:

$$\begin{aligned} \varrho(\delta) &= M[l(W, W^*)] \\ &= \iiint \iiint l(w, w^*)f(s)f_1(z|s)f_2(w|s)\delta(w^*|z)dsdzdwdw^* \end{aligned} \quad (63)$$

where the integration is extended over the domain of all possible values of the signals  $S$ ,  $Z$ ,  $W$ ,  $W^*$ .

Noting that

$$f(s)f_1(z|s) = p(z)p_1(s|z) \quad (64)$$

where

$$p(z) = \int f(s)f_1(z|s)ds \quad (65)$$

is the unconditional probability density of the input  $Z$ , and  $p_1(s|z)$  is the conditional probability density of the signal  $S$  with respect to  $Z$ , we may write eqn (63) in the form

$$\varrho(\delta) = \int p(z)dz \int \varrho(w^*, z)\delta(w^*|z)dw^* \quad (66)$$

where

$$\varrho(w^*, z) = \iint l(w, w^*)p_1(s|z)f_2(w|s)dsdw \quad (67)$$

The function  $\varrho(w^*, z)$  represents the *average conditional risk*, corresponding to definite realizations  $w^*$ ,  $z$  of signals  $W^*$ ,  $Z$ .

It is evident, that if for each realization  $z$  of the input  $Z$  there exists a unique value  $\hat{w}^*$  of the variable  $w^*$ , for which

$$\varrho(\hat{w}^*, z) = \inf_{w^*} \varrho(w^*, z) = \min_{w^*} \varrho(w^*, z) \quad (68)$$

then the Bayesian decision minimizing the average risk  $\varrho(\delta)$  is the Dirac  $\delta$ -function (i.e. the probability distribution condensed in a single point  $\hat{w}^*$ ):

$$\delta(w^*|z) = \delta(w^* - \hat{w}^*) \quad (69)$$

The value  $\hat{w}^*$  is evidently dependent on  $z$ :

$$\hat{w}^* = Az \quad (70)$$

where  $A$  is a definite operator.

Thus the minimizing of the average risk yields the unique deterministic operator of the optimal system

$$W^* = AZ \quad (71)$$

provided that for each realization  $z$  of the signal  $Z$  there exists a unique value  $\hat{w}^*$  satisfying (68).

In the case where the unique  $\hat{w}^* = Az$  satisfying (68) doesn't exist, the optimal system operator (optimal algorithm of information processing) may be random.

The above propositions of the statistical decision theory can evidently be directly applied only when the signals  $S$ ,  $Z$ ,  $W$  and  $W^*$  are finite-dimensional vectors. Hence they are immediately applicable to problems of sampled-data systems design only. It is certainly possible to rewrite them for the case where  $S$ ,  $Z$ ,  $W$  and  $W^*$  are random elements in any spaces, particularly, functional, using corresponding probability measures instead of probability densities. But the direct application of the theory would be impracticable in this case. Therefore another way of the extension of the above results to continuous systems is preferable, that is, the passing to the limit when the number of dimensions becomes infinite.

The limiting process in the preceding formulas, when the number of dimensions becomes infinite, may be accomplished in two different ways.

The first way is to replace all functions of continuous arguments by their values at sampled points and then to increase indefinitely the number of these points, all distances between the consecutive points becoming zero. This way has been used by many authors, particularly by Laning and Battin<sup>2</sup>, Sect. 8-6 and Middleton<sup>63</sup>. But this approach leads to considerable mathematical difficulties, owing to which the way of proving rigorously the existence of all limits encountered has not yet been found. Hence the passing to the limit from the sampled-data case to the continuous one remains, in the meanwhile, purely heuristic and is not strictly founded.

Another way is to express all random functions of continuous arguments in terms of random parameters and to pass to the limit from a finite set of random parameters to a denumerable one. This approach enabled the rigorous theory to be developed for determining operators of deterministic optimal systems under rather general conditions<sup>64-70</sup>, the random functions being expressed in terms of random parameters with the aid of canonical expansions of the form of eqn (3).

We shall confine ourselves here to outlining the method derived<sup>64-70</sup> for the simplest case, which covers nevertheless, a sufficiently wide variety of practical problems.

In many practical problems the input signal  $S$  doesn't contain irregular components and is the known function (non-linear in general) of the parameters  $U_1, \dots, U_N$  and of time  $t$ , and the noise is additive and normally distributed. The input  $Z$  and the desired output  $W$  are expressed by

$$Z(t) = \varphi(t, U) + X(t), \quad W(t) = \psi(t, U) \quad (72)$$

in such cases,  $U$  being the vector signal parameter with components  $U_1, \dots, U_N$ ,  $\varphi$  and  $\psi$  are given functions of  $t$  and  $U$ , and  $X(t)$  is a normally distributed noise independent of the vector  $U$ . Then the method derived in<sup>64-70</sup> yields the following general result (see<sup>12, Sect. 16.1</sup>): the optimum is attained by such a system which, receiving at the input  $Z$ , is processing output by minimizing the integral

$$I(W^*, Z) = \int_{-\infty}^{\infty} I(\psi(t, u), W^*) \exp\{L(u)Z(\tau) - \frac{1}{2}\beta(u)\} f(u) du \quad (73)$$

$f(u)$  being the joint probability density of all random components of  $U$  ( $f(u) \equiv 1$ , when all the components of  $U$  are unknown non-random quantities which can assume any values),  $L(u)$  is the linear operator

$$L(u)Z(\tau) = \int_{t-T}^t g(t, \tau, u) Z(\tau) d\tau \quad (74)$$

whose vector of weighting functions  $g(t, \tau, u)$  is defined by simultaneous linear integral equations of the form of eqn (56)

$$\int_{t-T}^t g(t, \tau, u) K_x(\tau, \sigma) d\tau = \varphi(\sigma, u) \quad (t-T \leq \sigma \leq t) \quad (75)$$

and the function  $\beta(u)$  is defined by

$$\beta(u) = L(u)\varphi(\tau, u) = \int_{t-T}^t g(t, \tau, u) \varphi(\tau, u) d\tau \quad (76)$$

The above relations determine the algorithms of optimal systems of various types, depending on the choice of the loss function (or functional) and the way in which function  $\varphi$  is related to  $\varphi$ . In particular, they yield the general solution of signal detection problem for signals depending on a finite number of unknown parameters<sup>1, Sect. 141, 142</sup>. The procedure outlined above proves also applicable to certain cases of infinite number of signal parameters  $U_1$ ,<sup>1, Sect. 144, and 70</sup>.

In the case, where  $Z$  and  $W$  are linear functions of signal parameters  $U_1, \dots, U_N$ , the probability density  $f(u)$  is normal (or  $f(u) \equiv 1$ ), and the loss-function (functional) is dependent on the system error  $W^* - W$  only, the optimal system, processing in accordance with eqns (73)–(76) is linear<sup>1, Sect. 143, 145</sup>.

The above method is easily generalized and is also apt for much more complicated situations. In particular, it is applicable, when the input signal  $Z$  contains an additive irregular component, or when the irregular part of the input signal and the noise are not additive, provided that the input  $Z$  is reducible to the form of the first eqn (72) by a known non-linear transform<sup>1, Sect. 144</sup>. This general method allows the solution of various problems of detection, discerning and extraction of signals arising in automatic control and radio engineering, and also of analogous problems in other branches of sciences and engineering. This method is applicable to single-dimensional systems and to multi-dimensional systems with an arbitrary number of inputs and outputs as well, the functions  $Z$ ,  $W$  and  $W^*$  being vectors in the latter case. Applying the method of canonical expansions to solve eqn (75) the above method is also applicable when  $Z$ ,  $W$  and  $W^*$  are functions not only of time,

but also of other arguments, for example, of coordinates of points in the space.

Optimal systems algorithms defined by the above method may be used in real automatic systems containing digital computers (sufficiently high-speed). In particular, these algorithms may be used in the design of 'self-learning' systems automatically studying the statistics of inputs and processing the control programme accordingly<sup>66</sup>.

The mathematical basis for simulating the processes of learning and experience storing in automatic systems has been developed with the aid of the statistical decision theory by the group of Czechoslovak mathematicians with the late A. Shpachek at the head<sup>71-74</sup>.

The peculiarity of feedback systems is that the input of the part of a system involved in a feedback loop is dependent on its output. The dynamical properties of any system (or any component of a system) are characterized in the general case by the conditional probability density of output for a given input. If one of the inputs of a certain part of the system is dependent on its output, it is impossible to define the conditional probability density of output with respect to input for this part of the system. Consequently it is impossible to define directly the conditional probability density of output for a given input for the system component involved in a feedback loop. Hence the statistical decision theory can be applied in the design of systems components involved in feedback loops, only if some additional assumptions are made, or an iterative process must be used. In particular, the statistical decision theory may be directly used for sampled-data systems with delayed feedbacks, the inputs of which at any time are dependent only on the values of respective outputs at past instants. The statistical decision theory yields in this case multi-stage optimal decisions defining the optimal algorithms of sampled-data systems with delayed feedbacks. It is by this method that Feldbaum has developed the so called dual control theory, using the statistical decision theory<sup>75-78</sup>.

## Conclusions

We see, that statistical methods in automatic control are now well developed and represent a powerful tool for the design and study of automatic systems and for the solution of automatic control problems. Voluminous literature has been involved in this field. We have considered here only the most important basic problems of the statistical theory of automatic control, and many important works done in various countries by scientists of various nationalities have been mentioned. Nevertheless all papers mentioned here represent only a small part of scientific papers in the field of statistical methods in automatic control.

In spite of the progress achieved, many statistical problems of the automatic control theory are not solved yet. Many problems are waiting for their statement and solution.

The basic problem of further development of the statistical theory in the near future is undoubtedly the creation of the statistical theory of complex systems and of methods for controlling complex systems having a statistical nature.

Automation is now expanding to a constantly increasing number of fields of human activity. In particular, the works in the field of automatization of diagnostics in medicine and in the field of application of methods of automatic control theory to biological problems and to problems of other branches of sciences



as well have been started. Statistical methods of automatic control theory will undoubtedly be useful in these fields and so will contribute to the progress of mankind.

## References

- <sup>1</sup> PUGACHEV, V. S. *Theory of random functions and its application to automatic control problems*. 1st edn 1957. Moscow-Leningrad; Gostekhizdat; 2nd edn 1960. Moscow-Leningrad; Fizmatgiz; 3rd edn 1962. Moscow-Leningrad; Fizmatgiz
- <sup>2</sup> LANING, J. H. and BATTIN, R. H. *Random processes in automatic control*. 1956. New York; McGraw-Hill
- <sup>3</sup> SOLODOV, A. V. *Automat. Telemekh.*, XIX, No. 4 (1958) 312-324
- <sup>4</sup> NEKOLNY, YA. and BENES, I. *Proc. 1st I.F.A.C. Congr., Moscow*, 1960. II. 1961. London; Butterworths
- <sup>5</sup> PUGACHEV, V. S. *Trans. Inst. Radio Engrs*, CT-7, No. 4 (1960) 506-512
- <sup>6</sup> BOOTON, R. C. *Proc. Symp. Nonlinear Circuit Analysis*, 2 (1953) 369-391
- <sup>7</sup> KAZAKOV, I. E. *Automat. Telemekh.*, XVII, No. 5 (1956) 385-409
- <sup>8</sup> KAZAKOV, I. E. and DOSTUPOV, B. G. *Statistical dynamics of nonlinear automatic systems*. 1962 (in Russian). Moscow-Leningrad; Fizmatgiz
- <sup>9</sup> KAZAKOV, I. E. *Proc. 1st I.F.A.C. Congr., Moscow*, 1960. II. 1961. London; Butterworths
- <sup>10</sup> PUPKOV, K. A. *Automat. Telemekh.*, XXI, No. 2 (1960) 191-200
- <sup>11</sup> PERVOZVANSKY, A. A. *Random processes in nonlinear automatic systems*. 1962 (in Russian). Moscow-Leningrad; Fizmatgiz
- <sup>12</sup> *Fundamentals of Automatic Control* (ed. V. S. PUGACHEV) 1963 (in Russian). Moscow-Leningrad; Fizmatgiz
- <sup>13</sup> SOMMERVILLE, M. J. and ATHERTON, D. P. *Proc. Inst. elect. Engrs* 105, No. 8 (1958) 537-549
- <sup>14</sup> POPOV, E. P. and PALTOV, I. P. *Approximate methods of studying nonlinear systems*. 1960 (in Russian). Moscow-Leningrad; Fizmatgiz
- <sup>15</sup> SAWARAGI, Y. and SUNAHARA, Y. *Tech. Rep. Engng. Res. Inst. Kyoto Univ*, VIII, No. 5 (1958) 95-126; No. 10, 195-218; IX, No. 5 (1959) 77-96; No. 12, 209-222; X, No. 1 (1960) 1-70
- <sup>16</sup> SAWARAGI, Y. and SUGAI, N. *Tech. Rep. Engng. Res. Inst. Kyoto Univ.*, IX, No. 8 (1959) 129-140
- <sup>17</sup> SAWARAGI, Y. and SUGAI, N. *Tech. Rep. Engng. Res. Inst. Kyoto Univ.*, IX, No. 9 (1959) 141-157
- <sup>18</sup> SAWARAGI, Y. and SUNAHARA, Y. *Tech. Rep. Engng. Res. Inst. Kyoto Univ.*, X, No. 4 (1960) 43-58
- <sup>19</sup> SAWARAGI, Y., SUNAHARA, Y. and NAKAMIZO, T. *Tech. Rep. Engng. Res. Inst. Kyoto Univ.*, XI, No. 1 (1961) 1-18
- <sup>20</sup> DOSTUPOV, B. G. *Automat. Telemekh.*, XVIII, No. 7 (1957) 999 to 1009
- <sup>21</sup> DOSTUPOV, B. G. and PUGACHEV, V. S. *Automat. Telemekh.*, XVIII, No. 7 (1957) 620-630
- <sup>22</sup> FELDBAUM, A. A. *Automat. Telemekh.*, XX, No. 8 (1959) 1056 to 1070
- <sup>23</sup> FELDBAUM, A. A. *Automat. Telemekh.*, XXI, No. 2 (1960) 167 to 179
- <sup>24</sup> PERVOZVANSKY, A. A. *Izv. Akad. Nauk SSSR, OTN, Energetika automat.*, No. 3 (1960) 64-72
- <sup>25</sup> PERVOZVANSKY, A. A. *Izv. Akad. Nauk SSSR, OTN, Energetika automat.*, No. 5 (1960) 187-195
- <sup>26</sup> KRASOVSKY, A. A. *Izv. Akad. Nauk SSSR, OTN, Energetika automat.*, No. 3 (1960) 37-45
- <sup>27</sup> KRASOVSKY, A. A. *Izv. Akad. Nauk SSSR, OTN, Energetika automat.*, No. 4 (1960) 121-129
- <sup>28</sup> KRASOVSKY, A. A. *Proc. 1st I.F.A.C. Congr., Moscow*, 1960. II. 1961. London; Butterworths
- <sup>29</sup> PUGACHEV, V. S. *Izv. Akad. Nauk SSSR, Math. ser.*, 17, No. 5 (1953) 401-420
- <sup>30</sup> ZADEH, L. A. *J. appl. Phys.* 24, No. 4 (1953) 396-404
- <sup>31</sup> ZADEH, L. A. *J. Franklin Inst.* 255, No. 5 (1953) 387-408
- <sup>32</sup> LUBBOCK, J. K. *Proc. Inst. elect. Engrs* 107, Part C, No. 11 (1960) 60-74
- <sup>33</sup> LUBBOCK, J. K. *Proc. 1st I.F.A.C. Congr., Moscow*, 1960, vol. II, 1961. London; Butterworths
- <sup>34</sup> PRASAD, T. *Proc. 1st I.F.A.C. Congr., Moscow*, 1960, vol. I, 1961. London; Butterworths
- <sup>35</sup> WIENER, N. *Nonlinear Problems in Random Theory*. 1958. New York; Wiley
- <sup>36</sup> ANDRONOV, A. A., VITT, A. A. and PONTRYAGIN, L. S. *J. exp. theor. Phys.* 3, No. 3 (1933) 165-180
- <sup>37</sup> BARRETT, J. F. *Proc. 1st I.F.A.C. Congr., Moscow*, 1960. Vol. II, 1961. London; Butterworths
- <sup>38</sup> PUGACHEV, V. S. *Proc. 1st I.F.A.C. Congr., Moscow*, 1960. Vol. II, 1961. London; Butterworths
- <sup>39</sup> PUGACHEV, V. S. *Izv. Akad. Nauk SSSR, OTN, Energetika automat.*, No. 3 (1961) 46-57
- <sup>40</sup> HAZEN, E. M. *Teoria veroyatnostey i ee primeneniya*, VI, No. 1 (1961) 130-137
- <sup>41</sup> HAZEN, E. M. *Izv. Akad. Nauk SSSR, OTN, Energetika automat.*, No. 3 (1961) 58-72
- <sup>42</sup> GOODMAN, T. P. and RESWICK, J. B. *Trans. Amer. Soc. mech. Engrs*, 78, No. 2 (1956) 259-271
- <sup>43</sup> LEONOV, YU. P. and LIPATOV, L. N. *Automat. Telemekh.*, XX, No. 9 (1959) 1289-1301
- <sup>44</sup> ANDREEV, N. I. *Automat. Telemekh.*, XVIII, No. 7 (1957) 615-619
- <sup>45</sup> ANDREEV, N. I. *Automat. Telemekh.*, IX, No. 12 (1958) 1077-1090
- <sup>46</sup> ANDREEV, N. I. *Automat. Telemekh.*, XX, No. 7 (1959) 833-838
- <sup>47</sup> PUGACHEV, V. S. *Automat. Telemekh.*, XVIII, No. 11 (1957) 971 to 984
- <sup>48</sup> PUGACHEV, V. S. *Trudy VI Vsesoyuznogo soveshchania po teorii veroyatnostey i matematicheskoy statistike*, 1960. Gos. izdat. polit. i nauchn. literatury Litovskoy SSR, Vilnius, 1962
- <sup>49</sup> WIENER, N. *Extrapolation, Interpolation and Smoothing of Stationary Time Series*. 1949. New York; Wiley
- <sup>50</sup> BOOTON, R. C. *Proc. Inst. Radio Engrs N.Y.*, 40, No. 8 (1952) 977-981
- <sup>51</sup> ZADEH, L. A. and RAGAZZINI, J. R. *J. appl. Phys.* 21, No. 7 (1950) 645-655
- <sup>52</sup> KALMAN, R. E. *Trans. Amer. Soc. mech. Engrs* 82 (1960) 35-45
- <sup>53</sup> KALMAN, R. E. and BUCY, R. S. *Trans. Amer. Soc. mech. Engrs*, Paper No. 60-JAC-12, 1960
- <sup>54</sup> LETOV, A. M. *Automat. Telemekh.*, No. 4 (1960) 436-441; No. 5, 561-568; No. 6, 661-665; XXII, No. 4 (1961) 425-435; XXIII, No. 11 (1962) 1405-1413
- <sup>55</sup> PONTRYAGIN, L. S., BOLTYANSKY, V. G., GAMKRELIDZE, R. V., MISHCHENKO, E. F. *Mathematical Theory of Optimal Processes*. 1961 (in Russian). Moscow-Leningrad; Fizmatgiz
- <sup>56</sup> KRASOVSKY, N. N. *Appl. Math. Mech., Leningr*, XXIV, No. 1 (1960) 64-79; XXV, No. 5 (1961) 806-817
- <sup>57</sup> KRASOVSKY, N. N. and LIDSKY, E. A. *Automat. Telemekh.*, XXII, No. 9 (1961) 1145-1150; No. 10, 1273-1278; No. 11, 1425-1431
- <sup>58</sup> ZAITSEV, A. G. *Automat. Telemekh.*, XXIV, No. 2 (1963) 143-150; No. 4, 447-454
- <sup>59</sup> PUGACHEV, V. S. *Automat. Telemekh.*, XVII, No. 6 (1956) 489-499
- <sup>60</sup> WALD, A. *Statistical Decision Functions*. 1950. New York; Wiley
- <sup>61</sup> VAN METER, D. and MIDDLETON, D. *Trans. Inst. Radio Engrs*, IT-4 (1954) 119-145
- <sup>62</sup> MIDDLETON, D. and VAN METER, D. *Journ. Soc. Indust. Appl. Math.*, 3 (1955) 192-253; 4 (1956) 86-119
- <sup>63</sup> MIDDLETON, D. *An introduction to Statistical Communication Theory*. 1960. New York; McGraw-Hill
- <sup>64</sup> PUGACHEV, V. S. *Automat. Telemekh.*, XIX, No. 6 (1958) 519-539
- <sup>65</sup> PUGACHEV, V. S. *Trans. 2 Prague Conf. on Inf. Theory* (held in 1959). 1960. Publ. House Czech. Acad. Sci.



- <sup>66</sup> PUGACHEV, V. S. *Proc. 1st I.F.A.C. Congr., Moscow, 1960*. Vol. II, 1961. London; Butterworths
- <sup>67</sup> PUGACHEV, V. S. *Izv. Akad. Nauk SSSR, OTN, Energetika avtomat.*, No. 2 (1960) 83-97
- <sup>68</sup> PUGACHEV, V. S. *Trans. Inst. Radio Engrs*, v. IT-6, No. 1 (1960) 4-7
- <sup>69</sup> PUGACHEV, V. S. *Trans. Inst. Radio Engrs*, v. CT-7, No. 4 (1960) 491-505
- <sup>70</sup> PUGACHEV, V. S. *Izv. Akad. Nauk SSSR, OTN, Energetika avtomat.*, No. 5 (1961) 123-135
- <sup>71</sup> SHPACHEK, A. *Trans. 1 Prague Conf. on Inf. Theory* (held in 1956) 1957. Publ. House Czech. Acad. Sci.
- <sup>72</sup> DRIML, N. and SHPACHEK, A. *Trans. 1 Prague Conf. on Inf. Theory* (held in 1956) 1957. Publ. House Czech. Acad. Sci.
- <sup>73</sup> HANSH, O. *Trans. 1 Prague Conf. on Inf. Theory* (held in 1956) 1957. Publ. House Czech. Acad. Sci.
- <sup>74</sup> WINKELBAUER, K. *Trans. 1 Prague Conf. on Inf. Theory* (held in 1956) 1957. Publ. House Czech. Acad. Sci.
- <sup>75</sup> FELDBAUM, A. A. *Automat. Telemekh.*, XXI, No. 9 (1960) 1240 to 1249; No. 11, 1453-1464; XXII, No. 1 (1961) 3-16; No. 2, 129 to 142
- <sup>76</sup> FELDBAUM, A. A. *Izv. Akad. Nauk SSSR, OTN, Energetika avtomat.*, No. 4 (1961) 107-119
- <sup>77</sup> FELDBAUM, A. A. *Automat. Telemekh.*, XXIII, No. 8 (1962) 993 to 1007
- <sup>78</sup> FELDBAUM, A. A. *Izv. Akad. Nauk SSSR, OTN, Technicheskaya Kibernetika*, No. 1 (1963) 13-25

# A New Method to Derive the Describing Function of Certain Non-linear Transfer Systems

R. LAUBER

## Summary

Automatic control system components may often be described by non-linear differential equations with products or powers of the system variables. Besides the examples cited below, every adaptive control system contains products of variables. Another important class of control systems are those with periodic coefficients. To determine the stability of control systems of this kind, one is interested in the describing function of its components.

To derive the describing function of such non-linear systems a method is presented, which is very similar to the well-known procedures to calculate frequency-responses of linear systems. It yields a system of algebraic equations with complex variables which is an implicit form of the describing function. The derivation of these equations is shown in a general manner.

To demonstrate the application of the method, several examples are considered. The first is a transfer system with a product of the input and output variable. Using all higher harmonics in the output, a describing function is derived which is an infinite continued fraction. A practical case of a system of this form is a nuclear reactor. The results for this second example are briefly demonstrated. In a third example the method is applied to a non-linear system with a square of the output variable.

## Sommaire

En étudiant des asservissements non-linéaires on est conduit souvent à des systèmes physiques représentés non pas par une caractéristique, mais par des équations différentielles contenant des produits ou des puissances des variables. Outre les exemples cités ce sont surtout les asservissements adaptatifs qui contiennent des produits des variables. Pour déterminer la stabilité des systèmes de ce type une analyse sinusoidale (au lieu d'une analyse temporelle) est suggérée ici, très semblable à la méthode fréquentielle qu'on emploie dans le cas des systèmes linéaires. On obtient donc un système d'équations complexes contenant la fonction de transfert sous forme implicite.

L'application de la méthode suggérée est exposée sur plusieurs exemples. Premièrement un système simple caractérisé par une équation différentielle contenant un produit du signal appliqué  $x_e(t)$  et de la réponse indicielle  $x_a(t)$  est calculé. La fonction de transfert implicite de ce système contenant toutes les harmoniques de la réponse forme des équations linéaires complexes, dont la solution explicite donne une fraction continue infinie.

A titre d'exemple pratique et important d'un système de ce type, le réacteur nucléaire est brièvement traité. Le troisième exemple est un système physique représenté par une équation différentielle non-linéaire contenant des puissances des variables.

## Zusammenfassung

Die Bestandteile eines automatischen Regelsystems können oft durch nichtlineare Differentialgleichungen beschrieben werden, welche Produkte oder Potenzen der Systemveränderlichen enthalten. Außer in den am Ende angeführten Beispielen treten in jedem adaptierenden System Produkte der Veränderlichen auf. Um die Stabilität solcher Anordnungen zu bestimmen, benötigt man die Beschreibungsfunktion der Bausteine.

Zur Ermittlung der Beschreibungsfunktion derartiger nichtlinearer Systeme wird ein Verfahren vorgeschlagen, welches der bekannten

Bestimmung der Frequenzgänge linearer Systeme ähnlich ist. Es liefert ein allgemeingültiges System algebraischer Gleichungen mit komplexen Veränderlichen, welches in impliziter Form die Beschreibungsfunktion darstellt.

Der Veranschaulichung dienen mehrere Beispiele. Das erste ist ein Übertragungssystem, in dem das Produkt der Eingangs- und Ausgangsgröße auftritt. Unter Benutzung aller Oberschwingungen der Ausgangsgröße ergibt sich die Beschreibungsfunktion als unendlicher Kettenbruch. Ein der Praxis entnommenes Beispiel, für welches einige Ergebnisse mitgeteilt werden, ist der Kernreaktor. Zum Abschluß wird die Anwendung des Verfahrens auf ein nichtlineares System gezeigt, in dem das Quadrat der Ausgangsgröße von Bedeutung ist.

## Introduction

Instabilities in feedback control systems will mostly cause oscillations. Linear systems are able to oscillate sinusoidally; non-linear systems may exhibit sustained periodic oscillations. Frequently these will be nearly sinusoidal, which means that there exists a fundamental component predominant to higher harmonics. If this holds, the stability of the non-linear control system can be investigated by considering only one or at least very few harmonics. The well-known frequency-response methods are then applicable, provided the describing function is used in the same way as the frequency response function of a linear component.

In its usual form the describing function describes the behaviour of a non-linear element for harmonic inputs. Assuming a pure harmonic variable

$$x_e(t) = \hat{x}_{e1} \sin \omega t \quad (1)$$

at the input which may cause a periodic output (see Figure 1)

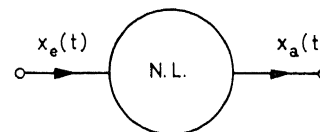


Figure 1. Non-linear system

$$x_a(t) = \hat{x}_{a1} \sin(\omega t + \phi_1) + \hat{x}_{a2} \sin(2\omega t + \phi_2) + \dots \quad (2)$$

the describing function is defined as the ratio

$$N = \frac{\hat{x}_{a1} e^{j\phi_1}}{\hat{x}_{e1}} \quad (3)$$

This definition can be generalized by assuming more than one harmonic at the input of the non-linear component

$$x_e(t) = \sum_{v=0}^n \hat{x}_{ev} \sin v\omega t \quad (4)$$

producing a periodic output

$$x_a(t) = \sum_{v=0}^{\infty} \hat{x}_{av} \sin(v\omega t + \phi_v) \quad (5)$$

If only the fundamentals of input and output are again compared the same definition (3) as shown above is obtained.

It should be mentioned though, that the consideration of higher harmonics in the input function  $x_e(t)$  will be necessary only in very few cases; on the other hand, for many non-linear systems a term of order zero, the d.c. component  $x_{e0}$ , must be taken into account.

The calculation of the describing function is done usually in the time domain by a Fourier transform of the periodic output function  $x_a(t)$ . This method is quite effective for the most important non-linearities described<sup>1</sup> by a function  $x_a(t) = f[x_e(t)]$ .

Nevertheless, some non-linear control system components are described by an implicit function  $g[x_e(t), x_a(t)] = 0$  or by a differential equation  $D[x_e(t), \dot{x}_e(t), \dots, x_a(t), \dots] = 0$ . Here the explicit evaluation of the periodic output  $x_a(t)$  for harmonic input  $x_e(t)$  and thus the calculation of the describing function might be difficult. Methods are known to calculate describing functions for such systems in the time domain under the assumption that all higher harmonics in the output are zero, which may be the cause of serious errors.

Extending these approaches<sup>2,3</sup> a new method will be shown here for an evaluation in the frequency domain. Using complex methods analogous to those applied to derive frequency responses of linear systems, describing functions can frequently be derived in analytic form, if the non-linearities involved are products or powers of the system variables.

### General Method to Determine Describing Functions in the Frequency Domain

A non-linear system may be described by a differential equation

$$D[x_e(t), \dot{x}_e(t), \dots, x_a(t), \dot{x}_a(t), \dots] = 0 \quad (6)$$

The input variable  $x_e(t)$  may consist of one or several harmonics:

$$x_e(t) = \mathcal{R} \sum_{\mu=0}^m \bar{x}_{e\mu} e^{j\mu\omega t} \quad (7)$$

If the output will be periodic, it can be approximated by a sum of harmonics

$$x_a(t) = \mathcal{R} \sum_{v=0}^n \bar{x}_{av} e^{jv\omega t} \quad (8)$$

With these expressions (7) and (8) introduced, the differential equation (6) will be fulfilled the better, the more terms in sum (8) have been considered. On the other hand, a limited number of terms will give satisfactory results, as it is the core of the describing function approach to 'balance' harmonic terms rather than to solve exactly the differential equation. In the equation

derived by inserting the expressions (7) and (8), this principle corresponds to a comparison of coefficients of the terms of the same frequency.

Such a comparison is not possible directly, however, because according to the non-linearities involved there have to be formed non-linear functions of the real parts. On the premises made above, that only products or powers are allowed (which are the non-linearities occurring most frequently in practical control systems), this difficulty can be circumvented. It can be shown easily, that for a product of real parts the following formula holds:

$$\mathcal{R} \bar{x}_1 \mathcal{R} \bar{x}_2 = \frac{1}{2} \mathcal{R} (\bar{x}_1 \bar{x}_2 + \bar{x}_1 \bar{x}_2^*) \quad (10)$$

For powers of real parts one gets

$$(\mathcal{R} \bar{x})^k = \frac{1}{2^{k-1}} \mathcal{R} \left[ \sum_{v=0}^{k-1} \binom{k}{v} \bar{x}^{k-v} \bar{x}^{*v} \right] \quad \text{for } k \text{ odd}$$

$$= \frac{1}{2^{k-1}} \mathcal{R} \left[ \sum_{v=0}^{k-1} \binom{k}{v} \bar{x}^{k-v} \bar{x}^{*v} + \frac{1}{2} \binom{k}{k/2} \bar{x}^{k/2} \bar{x}^{*k/2} \right] \quad (11)$$

for  $k$  even.

Considering the formulae (10) and (11) and also the trivial relation

$$\mathcal{R} \bar{x}^* = \mathcal{R} \bar{x} \quad (12)$$

one can manipulate in eqn (9) the complex variables directly instead of the real parts. Thus a comparison of coefficients now possible yields a set of complex algebraic equations:

$$\begin{aligned} f_0(\bar{x}_{e\mu}, \bar{x}_{e\mu}^*, \bar{x}_{av}, \bar{x}_{av}^*) &= 0 \\ f_1(\bar{x}_{e\mu}, \bar{x}_{e\mu}^*, \bar{x}_{av}, \bar{x}_{av}^*) &= 0 \\ \vdots &\vdots \\ f_n(\bar{x}_{e\mu}, \bar{x}_{e\mu}^*, \bar{x}_{av}, \bar{x}_{av}^*) &= 0 \end{aligned} \quad \begin{aligned} \mu &= 0 \dots m \\ \nu &= 0 \dots n \end{aligned} \quad (13)$$

This set of equations now describes the behaviour of the non-linear system for an input according to (7). It thus represents the describing function in implicit form. It may be written explicitly according to the definition (3):

$$N = \frac{\bar{x}_{a1}}{\bar{x}_{e1}} \quad (14)$$

If the method presented above is applied to a linear system, all equations of the set (13) would prove to be identical and their solution would give the frequency response function. In the case of a non-linear system with products of the variables, a linear set of complex algebraic equations results. As will be shown in the first and second example, the solution of this set with an unlimited number of higher harmonics leads to a describing function, which has the form of a continued fraction. If the given differential equation contains powers of the variables, the

$$D \left[ \mathcal{R} \sum_{\mu=0}^m \bar{x}_{e\mu} e^{j\mu\omega t}, \mathcal{R} \sum_{\mu=0}^m j\mu\omega \bar{x}_{e\mu} e^{j\mu\omega t}, \dots, \mathcal{R} \sum_{v=0}^n \bar{x}_{av} e^{jv\omega t}, \mathcal{R} \sum_{v=0}^n jv\omega \bar{x}_{av} e^{jv\omega t}, \dots \right] = 0 \quad (9)$$

explicit solution of the set (13) proves to become increasingly difficult. In a third example it will be demonstrated that an explicit solution can be derived only if a limited number of higher harmonics of the output is taken into account.

### Examples

(1) The first simple example has been chosen so that the application of the general method can be illustrated in detail.

Consider a control system component characterized by a differential equation

$$\frac{dx_a(t)}{dt} = x_e(t) + x_e(t)x_a(t) \quad (15)$$

The non-linearity encountered is the product  $x_e(t)x_a(t)$ . The term 'non-linearity' seems to be incorrect, since the differential equation (15) is linear with a variable coefficient, if  $x_e(t)$  is a given time function. It is justified to speak of a non-linearity, however, if eqn (15) is used to describe the transfer behaviour of the system. Due to the product  $x_e x_a$ , the output  $x_a(t)$  depends non-linearly on the input  $x_e(t)$ .

According to the general method given above to calculate the describing function, the expressions (7) and (8) are introduced into the differential equation. For simplification the input function  $x_e(t)$  is assumed to be of the form

$$x_e(t) = \Re \sum_{\mu=0}^1 \bar{x}_{e\mu} e^{j\mu\omega t} = \Re (x_{e0} + \hat{x}_{e1} e^{j\omega t}) \quad (16)$$

The output  $x_a(t)$  may be expressed in a series of harmonics

$$x_a(t) = \Re \sum_{v=0}^n \bar{x}_{av} e^{jv\omega t} \quad (17)$$

With these terms introduced into the differential equation (15), and observing the formula (10) for a product of real parts, the following equation results:

$$\boxed{\text{Eqn (18)}}^*$$

Eqn (18) may be written in the form

$$\boxed{\text{Eqn (19)}}^\dagger$$

By comparing now the coefficients of the terms with the same frequencies, one gets the following set of complex equations:

$$x_{e0} + x_{e0}x_{a0} + \frac{1}{2} \Re \hat{x}_{e1} \bar{x}_{a1} = 0 \quad (20)$$

$$\begin{aligned} \bar{x}_{a1}(j\omega - x_{e0}) - \hat{x}_{e1} \left( 1 + x_{a0} + \frac{1}{2} \bar{x}_{a2} \right) &= 0 \\ \bar{x}_{a2}(2j\omega - x_{e0}) - \frac{\hat{x}_{e1}}{2} (\bar{x}_{a1} - \bar{x}_{a3}) &= 0 \\ \bar{x}_{a3}(3j\omega - x_{e0}) - \frac{\hat{x}_{e1}}{2} (\bar{x}_{a2} - \bar{x}_{a4}) &= 0 \\ \vdots &\vdots \\ \bar{x}_{an}(nj\omega - x_{e0}) - \frac{\hat{x}_{e1}}{2} (\bar{x}_{a(n-1)} - \bar{x}_{a(n+1)}) &= 0 \end{aligned} \quad (21)$$

The variables  $\bar{x}_{av}$  of the set (20) and (21) may be eliminated successively, giving the describing function

$$N = \frac{\bar{x}_{a1}}{\bar{x}_{e1}} = \frac{1 + x_{a0}}{j\omega - x_{e0} - \frac{1}{4} \hat{x}_{e1}^2 \frac{1}{2j\omega - x_{e0} - \frac{1}{4} \hat{x}_{e1}^2 \frac{1}{3j\omega - x_{e0} - \frac{1}{4} \hat{x}_{e1}^2 \frac{1}{j\omega \dots}}}} \quad (22)$$

In addition to (22), eqn (20) is valid for the constant components  $x_{a0}$  and  $x_{e0}$ .

It proves to be unnecessary in this case to limit the number of harmonics considered in the output function  $x_a(t)$ . For  $n \rightarrow \infty$  an infinite continued fraction results.

For a practical calculation of the describing function the general form of eqn (22) seems to be only of academic interest, since the results are quite satisfactory by taking into account only the first and second harmonics. If a digital computer is used to numerically calculate eqn (22), the consideration of higher harmonics is possible without any additional difficulties. In this case the describing function can be determined from eqn (22) with any degree of exactness. Furthermore, the general formula (22) gives insight into the behaviour of the system under consideration, showing the effect of higher harmonics on the results and making possible an estimate of the errors caused by omissions.

(2) As an important practical case of a component with a differential equation of the form discussed above, the nuclear reactor may be taken as a second example. In the region near a steady state power the kinetic equations<sup>4</sup> are:

$$\begin{aligned} \frac{1}{\beta} \frac{dx_a(t)}{dt} &= x_e(t) + x_e(t)x_a(t) - x_a(t) + \sum_{i=1}^6 a_i c_i(t) \\ \frac{dc_i(t)}{dt} &= \lambda_i [x_a(t) - c_i(t)] \quad i = 1 \dots 6 \end{aligned} \quad (23)$$

\* Eqn (18)

$$\Re \sum_{v=0}^n jv\omega \bar{x}_{av} e^{jv\omega t} = \Re (x_{e0} + \hat{x}_{e1} e^{j\omega t}) + \frac{1}{2} \Re \left[ \sum_{v=0}^n \bar{x}_{av} e^{jv\omega t} (x_{e0} + \hat{x}_{e1} e^{j\omega t}) + \sum_{v=0}^n \bar{x}_{av} e^{jv\omega t} (x_{e0} + \hat{x}_{e1} e^{-j\omega t}) \right] \quad (18)$$

† Eqn (19)

$$\Re \left[ \sum_{v=0}^n jv\omega \bar{x}_{av} e^{jv\omega t} - x_{e0} - \hat{x}_{e1} e^{j\omega t} - \hat{x}_{e0} \sum_{v=0}^n \bar{x}_{av} e^{jv\omega t} - \frac{1}{2} \hat{x}_{e1} \left( \sum_{v=0}^n \bar{x}_{av} e^{j\omega(v+1)t} + \sum_{v=0}^n \bar{x}_{av} e^{j\omega(v-1)t} \right) \right] = 0 \quad (19)$$

In this equation,  $l$ ,  $\beta$ ,  $a_i$  and  $\lambda_i$  are constants. The input  $x_e(t)$  (the so-called reactivity) will be assumed in the form of eqn (16). The periodic output  $x_a(t)$  (the relative deviation of the power from a steady-state value) and the variables  $c_i(t)$  (relative variation of the concentration of the  $i$ th delayed neutron group) are expressed as an infinite sum of harmonics:

$$x_a(t) = \mathcal{R} \sum_{v=0}^{\infty} \bar{x}_{av} e^{jv\omega t} \quad (24)$$

$$c_i(t) = \mathcal{R} \sum_{v=0}^{\infty} \bar{c}_{iv} e^{jv\omega t} \quad (25)$$

With the abbreviation

$$G(j\omega) = j\omega \left( \frac{l}{\beta} + \sum_{i=1}^6 \frac{a_i}{j\omega + \lambda_i} \right) \quad (26)$$

an insertion of these expressions (24) and (25) into the differential equation (23) conducted in the same way as in the first example yields the set of complex equations

$$\begin{aligned} x_{e0}(1 + x_{a0}) + \frac{1}{2} \mathcal{R} \hat{x}_{e1} \bar{x}_{a1} &= 0 \\ \bar{x}_{a1} [G(j\omega) - x_{e0}] - \hat{x}_{e1} \left( 1 + x_{a0} + \frac{1}{2} \bar{x}_{a2} \right) &= 0 \\ \bar{x}_{av} [G(jv\omega) - x_{e0}] - \frac{1}{2} \hat{x}_{e1} (\bar{x}_{a(v-1)} + \bar{x}_{a(v+1)}) &= 0 \end{aligned} \quad (27)$$

$v = 2 \dots \infty$

Elimination of the variables  $\bar{x}_{av}$  gives the describing function

$$\boxed{\text{Eqn (28)}}^*$$

It can easily be seen that the form of the describing function depends only on the type of the non-linearity. As in this second example, only the linear part of the equations is different from that of the first example, the resulting describing function (28) proves to be identical to the function (22), if only  $jv\omega$  is replaced by  $G(jv\omega)$ .

In Figures 2 and 3 the gain and phase of the describing function of the nuclear reactor are shown. In the case chosen here ( $x_{a0} = 0$ ), both the gain and phase are reduced with increasing amplitudes of the input reactivity. Hence the important conclusion follows, that stability will increase by the effects of the non-linearity compared to the linearized case (the linearized frequency response is the curve with zero amplitude)<sup>4</sup>.

(3) In a third example a transfer system with a power of the output variable is considered. A practical case of a system of this kind would be, for example, the heat transfer from a surface to a cooling medium. It can be described by a differential equation containing the fourth power of the output variable<sup>4, 5</sup>. For the sake of a simple illustration an investigation is made of a similar system with a square of the output variable.

It is assumed that the describing function of a system has to be calculated, characterized by the equation

$$\sum_{l=1}^m \frac{d^l x_a(t)}{dt^l} + x_a^2(t) = x_e(t) \quad (29)$$

Analogous to the preceding example the input variable is supposed to consist of a fundamental harmonic and a constant term  $x_{e0}$ .

The periodic variables are expressed in terms of their harmonic content:

$$\left. \begin{aligned} x_e(t) &= \mathcal{R} \sum_{\mu=0}^1 \bar{x}_{e\mu} e^{j\mu\omega t} \\ x_a(t) &= \mathcal{R} \sum_{v=0}^n \bar{x}_{av} e^{jv\omega t} \end{aligned} \right\} \quad (30)$$

With these expressions inserted into the differential equations (29) and applying the formula (11) which takes the form

$$(\mathcal{R} \bar{x})^2 = \frac{1}{2} \mathcal{R} (\bar{x}^2 + \bar{x} \bar{x}^*) \quad (31)$$

the following equation can be derived.

$$\begin{aligned} &\sum_{l=1}^m \mathcal{R} \sum_{v=0}^n (jv\omega)^l \bar{x}_{av} e^{jv\omega t} \\ &+ \frac{1}{2} \mathcal{R} \left[ \left( \sum_{v=0}^n \bar{x}_{av} e^{jv\omega t} \right)^2 + \left( \sum_{v=0}^n \bar{x}_{av} e^{jv\omega t} \right) \left( \sum_{v=0}^n \bar{x}_{av}^* e^{-jv\omega t} \right) \right] \\ &= \mathcal{R} \sum_{\mu=0}^1 \bar{x}_{e\mu} e^{j\mu\omega t} \end{aligned} \quad (32)$$

Comparison of the coefficients of the terms of identical frequencies yields the following set of algebraic equations:

$$\begin{aligned} x_{a0} + \frac{1}{2} \sum_{i=1}^n \hat{x}_{ai}^2 &= x_{e0} \\ \sum_{l=1}^m (j\omega)^l \bar{x}_{a1} + 2x_{a0} \bar{x}_{a1} + \sum_{i=1}^n \bar{x}_{ai}^* \bar{x}_{a(i+1)} &= \hat{x}_{e1} \\ \sum_{l=1}^m (2j\omega)^l \bar{x}_{a2} + 2x_{a0} \bar{x}_{a2} + \frac{1}{2} \left( \bar{x}_{a1}^2 + 2 \sum_{i=1}^n \bar{x}_{ai}^* \bar{x}_{a(i+2)} \right) &= 0 \\ \cdot &\cdot \cdot \cdot \cdot \\ \cdot &\cdot \cdot \cdot \cdot \cdot \\ \cdot &\cdot \cdot \cdot \cdot \cdot \\ \sum_{l=1}^m (nj\omega)^l \bar{x}_{an} + 2x_{a0} \bar{x}_{an} + \frac{1}{2} \sum_{i=1}^n (\bar{x}_{ai} \bar{x}_{a(n-i)} + 2\bar{x}_{ai}^* \bar{x}_{a(n+i)}) &= 0 \\ &n-i \geq 1 \end{aligned} \quad (33)$$

\* Eqn (28)

$$N = \frac{\bar{x}_{a1}}{\bar{x}_{e1}} = \frac{1 + x_{a0}}{G(j\omega) - x_{e0} - \frac{1}{4} \hat{x}_{e1}^2} \frac{1}{G(2j\omega) - x_{e0} - \frac{1}{4} \hat{x}_{e1}^2} \frac{1}{G(3j\omega) - x_{e0} - \frac{1}{4} \hat{x}_{e1}^2} \frac{1}{G(4j\omega) \dots} \quad (28)$$

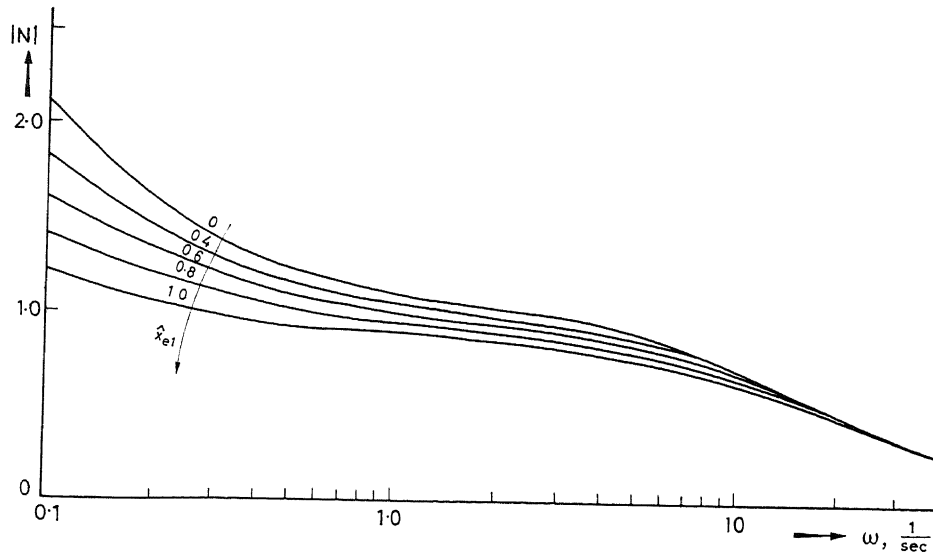


Figure 2. Amplitude of the describing function of a nuclear reactor according to eqn (28) ( $l/\beta = 0.1$  sec,  $x_{a0} \approx 0$ )

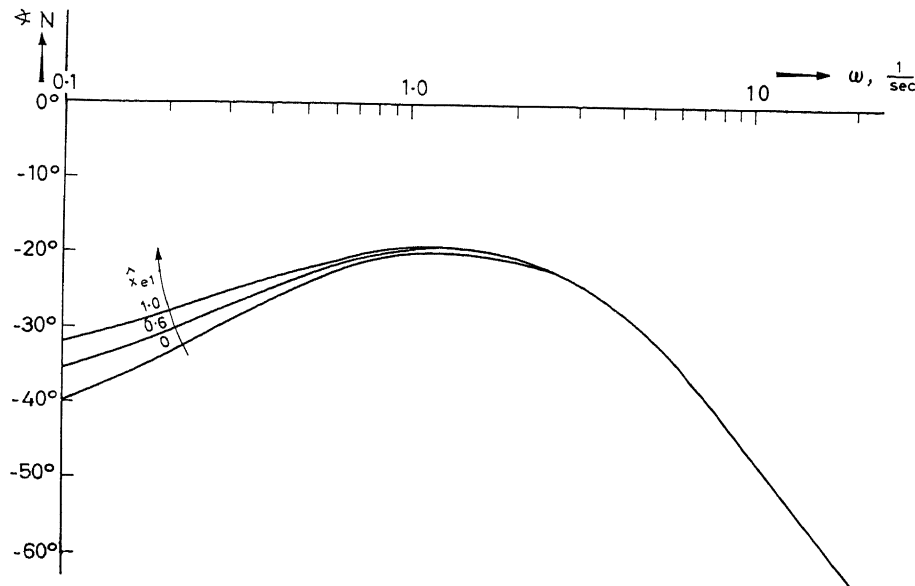


Figure 3. Phase of the describing function of a nuclear reactor

Contrary to the linear set of complex equations of the preceding examples two major difficulties arise: (a) Since the differential equation here was non-linear, the complex algebraic equations are now also non-linear. (b) The conjugate complex variables are contained explicitly in the resulting equations.

According to these complications, the analytic solution of the set of eqns (33) is possible only if a limited number of higher harmonics is allowed. Considering a second harmonic in addition to the d.c. component and the fundamental harmonic, the following formula can be derived from (33) for the describing function:

$$N = \frac{\bar{x}_{a1}}{\bar{x}_{e1}} = \frac{1}{\sum_{l=1}^m (j\omega)^l + 2x_{a0} + \frac{\hat{x}_{e1}^2}{2} \frac{|N|^2}{\sum_{l=1}^m (2j\omega)^l + 2x_{a0}}} \quad (34)$$

According to the first equation of (33) the d.c. component  $x_{a0}$  is a function of the input variables  $\hat{x}_{e1}$  and  $x_{e0}$ :

$$x_{a0} = \left( x_{e0} - \frac{1}{2} \hat{x}_{e1}^2 N^2 - \frac{1}{4} \hat{x}_{e1}^4 \left| \frac{N^2}{\sum_{l=1}^m (2j\omega)^l + 2x_{a0}} \right|^2 \right)^{\frac{1}{2}} \quad (35)$$

Eqns (34) and (35) can be solved numerically only by trial and error. It would thus be more practical to calculate  $\bar{x}_{a1}$  and the describing function  $N$  directly from the set of algebraic equations (33). The formulae (34) and (35) are only valuable to the extent that they show explicitly the influence of the second harmonic which is seen to depend on the square of the amplitude of the input. It can be neglected for very small amplitudes.

### Conclusions

The describing function of a control system component can effectively be evaluated in the frequency domain, if the non-

linearities involved consist of products or powers of the variables. A set of complex algebraic equations can be derived which contains the describing function in implicit form. For systems with other types of non-linearities than products or powers the functions of the real parts introduced could not be solved explicitly, which means that the method would not be applicable any more. In such cases the describing functions would have to be calculated in the time domain<sup>3, 2</sup>.

In the case of systems which contain only products of the input and output variables (which, for example, occur most frequently in adaptive control systems), or in other words which are characterized by linear differential equations with variable coefficients, the resulting complex algebraic equations will be linear. They can be solved explicitly for an unlimited number of higher harmonics in the output, and the describing function resulting has the form of an infinite continued fraction.

The complex algebraic equations will be non-linear, if a system characterized by a non-linear differential equation with powers of the variables is considered. Numerically the describing function can be calculated by iteration from these. If only a fundamental harmonic together with one or very few higher harmonics are taken into account, an explicit expression may be derived.

## References

- <sup>1</sup> LEONHARD, A. *Die selbsttätige Regelung*. 1957. Berlin; Springer-Verlag

- <sup>2</sup> MAGNUS, K. Über ein Verfahren zur Untersuchung nichtlinearer Schwingungs- und Regelungssysteme. *V.D.I.-Forsch.* 451  
<sup>3</sup> KLOTTER, K. Über den Gebrauch ersetzender Übertragungsfunktionen zur Untersuchung nichtlinearer Regelkreise. *Regelungstechnik Mod. Theorien und ihre Verwendbarkeit*. S. 160-165. 1957. Tagung Heidelberg; Oldenbourg-Verlag  
<sup>4</sup> LAUBER, R. Nichtlineare Stabilitätsuntersuchung von Reaktoren und Reaktor-Regelsystemen. *Atomkernenergie* 7 (1962), 95-101  
<sup>5</sup> GRÖBER, ERK, GRIGULL *Die Grundgesetze der Wärmeübertragung*. 1955. Berlin; Springer-Verlag

## Nomenclature

$x_e(t)$	Input variable of a transfer system
$x_a(t)$	Output variable of a transfer system
$\hat{x}$	Amplitude (maximal value) of a sinusoidal time function $x(t)$
$\bar{x} = \hat{x} e^{j\phi}$	Complex amplitude coefficient
$\bar{x}^* = \hat{x} e^{-j\phi}$	Conjugate complex amplitude coefficient
$\omega$	Frequency in rad/sec
$N$	Describing function
$\mathcal{R}$	Real part
Index 0	D. C. component
Index 1	1st harmonic (fundamental)
Index 2	2nd harmonic
etc.	

## DISCUSSION

W. W. SEIFERT, *Massachusetts Institute of Technology, Cambridge 39, Mass., U.S.A.*

The author is to be congratulated for a very lucid presentation of an interesting new method for deriving the describing function for systems in which the non-linearity appears in the form of products or powers of the variables. The method discussed carries out the evaluation completely in the frequency domain and offers the advantage of enabling one to consider a system input which consists of a number of harmonics rather than a simple sinusoid, and simultaneously permitting computation of the describing function when the output variable is described as a harmonic series. The real advantage of carrying out this computation in the frequency domain is that the whole process is reduced to an algebraic one, which, while the resultant equations may be very difficult to solve, is at least, straightforward to formulate. The difficulty of solution, however, follows directly from the basic complexity of the problem and certainly is no more severe in the proposed method than if the more usual time-domain approach were used.

On the other hand, for this method to be applicable, the non-linearity must be expressible in terms of products or powers of the variables. This is a relatively severe restriction but, in a sense, it is fortunate that the method applies to this type of system rather than some of those handled readily by time-domain methods. Systems involving products frequently lead to frequency-dependent describing functions whose derivation requires more ingenuity than those dependent on amplitude only. While the amount of actual labour involved is not reduced by Lauber's method, the process is at least well defined. Unfortunately, solution of the resultant equations may be too difficult to handle except by machine methods, with the result that in these cases the real advantage of the describing function method, namely that of being a simple to use engineering tool, is lost.

The method appears to offer no advantage whatsoever when the

output of the non-linearity is related to its input as a simple power, such as a square or a cube function. Here, it appears simpler to divide the problem into the linear portion and the non-linear portion for which the derivation of the describing function is essentially trivial, and handle the combination by the usual describing function method, than to develop one describing function for the overall system by Lauber's technique.

One problem which remains unanswered with either approach is that of determining the error which results when the output of the non-linearity is described by a few terms of a series, as is required if the amount of effort in obtaining numerical results is to be kept within bounds.

R. LAUBER, *in reply*

I thank Professor Seifert for his very profound comments, but regret that I cannot agree with the statement that the amount of labour is not reduced by my method. Making the calculation in the frequency domain, the variables are vectors containing both amplitude and phase. When the calculations are made in the time domain, amplitude and phase are treated separately so that the number of variables is doubled. This reduction of the number of variables by a factor of two might be very important in calculating describing functions of complicated systems, especially if higher harmonics are considered. I think that this is one of the advantages of my method.

I do agree with Professor Seifert that in those cases where the non-linear part of the system can easily be separated from the linear one, this should be done. The method presented is aimed at those problems where such a separation cannot easily be carried out.

To the question of the errors which result when considering only a few harmonics, no general answer can be given. I can only mention

that in the case of the nuclear reactor (the second example in my paper), an investigation to determine these errors has been carried out. It showed that a calculation without considering any higher harmonics in the output function leads to a small error in amplitude, but to quite an essential error in the phase of the describing function. A calculation without consideration of any higher harmonics leads to an error in phase  $\Delta \varphi \approx N = 10^\circ$  at the frequency  $\omega = 0.1 \text{ S}^{-1}$  (Figure 3 of the paper). If only the second harmonic is considered, the error at this point may be reduced to  $\Delta \varphi \approx N = 1^\circ$ . So in this example the consideration of one harmonic in addition to the fundamental harmonic leads to quite satisfactory results.

W. FILIPCZAK, *Institute of Nuclear Research, Swierk p. Otwock, Warsaw, Poland*

In your example concerning the nuclear reactor you reached some conclusions which are not clear to me. In Figure 2 of your paper you found that the absolute value of the describing function  $|N|$  is decreasing with the input amplitude  $\hat{X}_{e1}$  (the reactivity of the reactor). This does not agree with my view. I think the higher the disturbance in the reactivity the higher should be the value of  $|N|$ .

As I am an application engineer in the field of nuclear reactor control, I am interested to know the advantages of the describing function method over the usual simpler methods of analogue computer investigation. Is the application of the describing function method in this field only of purely academic interest? It may be that I omitted some important factors, but in the field of nuclear reactor control I do not see the urgent need of the describing function method. At the stage of its present development I am not sure that it can be easily and fruitfully applied to nuclear reactor control problems.

R. LAUBER, *in reply*

I thank Mr. Filipczak for his questions which are very important to an engineer working on practical problems in the field of nuclear reactor control. The answer to the question, if the amplitude of the describing function is always decreasing with increasing amplitudes of perturbation, is different for different examples. The important factor in this respect is the d.c. component  $X_{a0}$  at the output function  $X_a(t)$ , and this d.c. component  $X_{a0}$  in turn depends on the feedback of the reactor.

Figure 2 of my paper is valid for  $X_{a0} = 0$ . In the cases of pressurized water and boiling water reactors, the factor  $X_{a0}$  will not be zero, but so small, that still the amplitude of the describing function is decreasing with increasing amplitudes of perturbation. More information on this question can be found in Reference 4 of my paper.

The second question from Mr. Filipczak on the usefulness of the describing function method to reactor control problems cannot be answered in a general manner. At any rate, analogue computer methods are always necessary if transient responses have to be calculated, since the describing function method is not valid for the investigation of transients; but even for stability investigations it might be easier to use an analogue computer if special examples of control systems have to be considered.

On the other hand, the analogue computer study does not give general information, and if such information is wanted, an analytical method is necessary. A very instructive example can be found in Reference 4 of my paper. There it is concluded that for the case of boiling water reactors a linear treatment of the kinetic equations is always justified since stability will increase in the non-linear case with increasing amplitudes of perturbations. Such a general statement could never be made by interpreting diagrams computed with the aid of an analogue computer.



# The Describing Function Method Applied for the Investigation of Parametric Excited Oscillations

A. LEONHARD

## Summary

In control systems analysis there occur, occasionally, oscillatory loops including elements with multiplicative interaction. The investigation of such systems, using the describing function method in the usual manner, i.e. by neglecting the harmonics, may produce entirely wrong results. If the multiplier is excited by oscillations of various combination frequencies, it is quite possible that one harmonic may interact with a fundamental frequency component to form a new not-negligible component of fundamental frequency. In the simplest case a harmonic component  $y_2 = a_2 \sin 2\omega t$ , when multiplied with a component of fundamental frequency  $y_1 = a_1 \sin \omega t$  will produce an output of the form

$$y_s = \frac{1}{2} a_1 a_2 \cos \omega t - \frac{1}{2} a_1 a_2 \cos 3 \omega t$$

the fundamental component of which is a function of the harmonic excitation  $y_2$ . The harmonic components thus influence, often to an important degree, the fundamental component and the stability. This effect of interaction of various frequencies in the non-linear element is, of course, the basis for parametric oscillations.

When applying the describing function method to systems of this type—it is shown that this is advantageous—the harmonics must be taken into account in as much as they influence the fundamental frequency oscillation. Theoretically there is an infinite number of these harmonics. It is shown that satisfactory results can be obtained by including only two or three prevalent harmonics. The advantage of this method, as opposed to purely mathematical treatment, is the fact that it can be used with systems of any order and that it is a natural extension of the well-known linear control theory.

## Sommaire

Dans certains systèmes de commande, il existe des boucles instables comprenant des éléments à interactions multiplicatives. L'analyse de tels systèmes par la méthode de la fréquence fondamentale conduit quelquefois à des résultats erronés. Si le multiplieur est excité par une combinaison de fréquences, il est tout à fait possible que l'une des harmoniques agisse sur l'une des fréquences de base pour produire une composante non négligeable de cette dernière. Dans le cas le plus simple, la multiplication d'une harmonique

$$\begin{aligned} y_2 &= a_1 \sin 2\omega t && \text{par une fondamentale} \\ y_1 &= a_2 \sin \omega t && \text{produit une sortie de la forme:} \end{aligned}$$

$$y_s = \frac{1}{2} a_1 a_2 \cos \omega t - \frac{1}{2} a_1 a_2 \cos 3 \omega t$$

dont la composante fondamentale est une fonction de l'harmonique excitatrice  $y_2$ . Les harmoniques influencent ainsi, dans une large mesure, la composante fondamentale et la stabilité. Cet effet d'interaction de fréquences différentes dans l'élément non-linéaire est certainement à la base d'oscillations paramétriques.

Dans ce rapport, on présente une extension de la méthode de la fréquence fondamentale en tenant compte des harmoniques. Théoriquement, le nombre de ces dernières est infini. On montre que dans la pratique, il suffit de tenir compte de deux ou trois prédominantes. Une telle approximation peut être avantageusement utilisée avec des systèmes de n'importe quel ordre. Elle constitue une extension naturelle de la théorie linéaire.

## Zusammenfassung

In der Regelungstechnik kommen schwingungsfähige Gebilde mit Multiplikationsstellen vor. Untersucht man solche Kreise mit Hilfe

der Beschreibungsfunktion in ihrer herkömmlichen Form mit Vernachlässigung der Oberwellen, so kommt man unter Umständen zu vollkommen falschen Ergebnissen. An den Multiplikationsstellen treten verschiedene Frequenzkombinationen auf, und es kann durchaus der Fall auftreten, daß eine Oberwelle etwa zusammen mit der Grundwelle eine mit der Grundwelle gleichfrequente Schwingung als Ausgangsgröße liefert. Der einfachste Fall liegt vor, wenn die Oberschwingung  $y = a \sin 2\omega t$  multipliziert wird mit der Grundschiwingung  $b \sin \omega t$ , also  $a \sin 2\omega t \cdot b \sin \omega t = \frac{1}{2} a b \cos \omega t - \frac{1}{2} a b \cos 3\omega t$ . Am Ausgang erscheint also eine von der Oberschwingung abhängige, der Grundschiwingung gleichfrequente Schwingung. Die Oberschwingung beeinflusst also hier, unter Umständen sogar in erheblichem Ausmaß, die Grundschiwingung und damit die Selbsterregung. Parametrische Schwingungen, mit denen sich dieser Vortrag vor allem beschäftigen soll, kommen überhaupt erst durch solche Frequenzkombinationen zustande.

Will man also auch bei solchen Kreisen mit der Beschreibungsfunktion arbeiten und zeigt es sich, daß man dies mit Vorteil machen kann, so müssen die Oberwellen insoweit berücksichtigt werden, als sie die Grundwelle beeinflussen. Dies sind theoretisch im allgemeinen unendlich viele. Es läßt sich aber zeigen, daß schon bei einer Beschränkung auf nur zwei oder drei Oberwellen vielfach brauchbare Ergebnisse erzielt werden können. Die Vorteile dieser Behandlungsweise gegenüber anderen mathematischen Verfahren liegen einmal darin, daß die Zahl der Freiheitsgrade des Systems keine Rolle spielt, und zum anderen, daß sich die Methode an die bekannte lineare Regelungstheorie anlehnt.

## Introduction

The describing function method, sometimes called the harmonic balance concept, is already well known and provides a powerful means for the investigation of self-excited oscillations of electric machinery or for the stability analysis of non-linear control loops.

The basic assumption for the describing function concept is that the fundamental harmonic of an oscillation in a closed feedback loop is not influenced by the higher harmonics, which are in general damped by the lags of the loop anyway. All the higher harmonics will be neglected. Experience has taught that the method is a very powerful tool as long as the basic assumption is true, that the fundamental harmonic is predominant and not affected by the higher harmonics.

But there are technical systems where a multiplication of two variables is included in the dynamics<sup>1-3</sup>. Multiplication of variables in a dynamic system occurs frequently in all systems with time-varying parameters. Certain non-linear problems can be approached in the same way.

If one investigates such feedback loops with the aid of the usual describing function concept, in its conventional way, neglecting all the higher harmonics, one may obtain under certain conditions grossly incorrect results.

yields an output  $\alpha_3$  which gives at the output of  $B$ , after being multiplied by  $\gamma(\tau)$ , two oscillations with the frequencies  $\nu$  and  $5\nu$  designated by the symbols  $\beta_{1b}$  and  $\beta_5$ . The dropping of the fifth harmonic  $\beta_5$  is usually allowed.

The two oscillations  $\beta_{1b}$  and  $\beta_{1a}$  are added and one obtains

$$\begin{aligned}\beta_{1a} &= \alpha_1 B_{11} & \beta_{1b} &= \alpha_3 \cdot B_{31} \\ \alpha_3 &= \beta_5 F_3 & \alpha_1 &= (\beta_{1a} + \beta_{1b}) F_1\end{aligned}$$

The condition for the stability boundary is easily determined

$$\frac{\beta_{1a} + \beta_{1b}}{\varepsilon_1} = B_{11} F_1 + B_{13} F_3 B_{31} F_1 = 1 \quad (20)$$

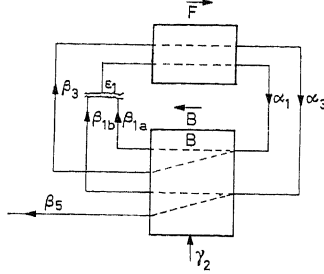


Figure 3. Corresponding block diagram to Figure 2 considering the third harmonic

According to the relation of eqn (9), the following equations are easily proved to be correct:

$$B_{11} = \frac{1}{2} \gamma (j \cos \psi - \sin \psi)$$

according to eqn (18).

$$B_{13} B_{31} = \frac{1}{4} \gamma^2 \quad (21)$$

The analogue eqn (17) for  $\mu = 2$  yields

$$F_3 = \frac{1}{\left(j 3 \frac{\mu}{2}\right)^2 + j 3 \frac{\mu}{2} \xi + 1} \quad (22)$$

So the final equations are obtained:

$$\begin{aligned}\frac{\beta_{1a} + \beta_{1b}}{\varepsilon_1} &= \frac{1}{2} \gamma \frac{j \cos \psi - \sin \psi}{\left(j \frac{\mu}{2}\right)^2 + j \frac{\mu}{2} \xi + 1} \\ + \frac{1}{4} \gamma^2 \frac{1}{\left[\left(j \frac{\mu}{2}\right)^2 + j \frac{\mu}{2} \xi + 1\right] \cdot \left[\left(j 3 \frac{\mu}{2}\right)^2 + j 3 \frac{\mu}{2} \xi + 1\right]} &= 1 \quad (23)\end{aligned}$$

If one performs the stability analysis with the more accurate eqn (23), compared with eqn (19), one obtains for  $\gamma$  with the specific values  $\mu = 2$  and  $4$  and  $\xi = 0.5$  practically the same values as one did with eqn (19). The influence of the third harmonic can therefore be neglected for these two problems. (That means for  $\mu \geq 2$ , the frequency of the sinusoidal parameter variation must be equal or larger than twice the resonance frequency.) One determines for the value  $\mu = 1$ , which means for  $\nu = \mu/2 = 0.5$ , with eqn (19), (without consideration of the

third harmonic)  $\gamma = 1.58$ . But one obtains for the stability analysis with the aid of eqn (23) (with consideration of the third harmonic oscillation) no real value for  $\gamma$ , which means there is no oscillation with the frequency  $\nu = 0.5$  possible for  $\mu = 1$ . As will be seen later, the fundamental harmonic for  $\mu = 1$  is no longer  $\nu = \mu/2$ , but  $\nu = \mu$ .

First, a second-order differential equation, like eqn (1), is investigated, but now it is assumed that the damping will vary harmonically. Eqn (24), according to eqn (16), is now

$$y'' + \xi_0 y' + y = y' \gamma \sin(\mu\tau + \psi) \quad (24)$$

( $\xi_0$  is the average value of the damping factor). Provided one again assumes steady-state sinusoidal oscillations, (the symbols are the same as in Figure 2)  $\alpha_1'$  is equal to  $d\alpha_1/d\tau_1$  and one obtains, corresponding to eqn (11) and eqn (9),

$$\beta_{1a} = \alpha_1 B_{11} = \alpha_1 \frac{\nu}{2} \gamma (\cos \psi + j \sin \psi) \quad (25)$$

or

$$\beta_3 = \alpha_1 B_{13} = \alpha_1 \frac{\nu}{2} \gamma e^{j\psi} \quad (26a)$$

$$\beta_{1b} = \alpha_3 B_{31} = \alpha_3 \left( -\frac{3\nu}{2} \gamma e^{-j\psi} \right) \quad (26b)$$

If one has first assumed the closed feedback loop to be open, the relation for self-excitation is again  $\beta_1/\varepsilon_1 = 1.0$ . Following this concept one finds the conditions for the determination of  $\psi$  and  $\gamma$  of eqn (24).

If the coefficient of  $y''$  is varying in the differential equation [it is assumed to be 1.0 in eqn (1)] periodically, one obtains, according to the eqns (25) and (26)

$$\beta_{1a} = \alpha_1 B_{11} = \alpha_1 \left[ -\frac{\nu^2}{2} \gamma (j \cos \psi - \sin \psi) \right] \quad (27)$$

$$\beta_3 = \alpha_1 B_{13} = \alpha_1 \left( +j \frac{\nu^2}{2} \gamma e^{j\psi} \right) \quad (28a)$$

$$\beta_{1b} = \alpha_3 B_{31} = \alpha_3 \left[ -j \frac{(3\nu)^2}{2} \gamma e^{-j\psi} \right] \quad (28b)$$

For the frequency ratio the assumption  $\nu/\mu = 0.5$  is made. With this the following oscillations with the frequencies besides  $\nu_1 = 0.5 \mu$  will occur  $\nu_3 = 1.5 \mu$ ;  $\nu_5 = 2.5 \mu$ ;  $\nu_7 = 3.5 \mu$ ;  $\nu_9 = 4.5 \mu$  and so on.

Which of these oscillations will be the most important depends on the resonance frequency. It should be noted here that the stability analysis can be performed with each of these oscillations.

The ratio of the fundamental harmonic to the varying parameter frequency of  $\nu/\mu = \frac{1}{2}$ , with the corresponding higher harmonics, is most likely to occur, but it is not the only possible one. Considering Figure 4, one will find, that a frequency ratio with the values  $\nu/\mu = 1$  is possible too, with the frequencies for the higher harmonics  $\nu_1 = 1.0 \mu$ ;  $\nu_0 = 0$ ;  $\nu_2 = 2 \mu$ ;  $\nu_3 = 3 \mu$  and so on.

The stability analysis for those problems is performed exactly in the same way as before with the value  $\nu/\mu = 0.5$ . All that remains to be done is to choose the proper values for

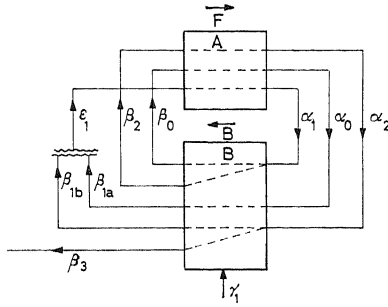


Figure 4. Block diagram for the calculation of the stability boundary  $\nu/\mu = 1.0$

the transfer functions  $F$  and  $B$ . According to Figure 4 one obtains for instance

$$\beta_2 = B_{12}\alpha_1$$

or

$$\beta_{1b} = B_{21}\alpha_2$$

and so on.

For an arbitrary system  $A$ , a self-excited oscillation can sustain either with one or other frequency ratio dependent on the frequency  $\mu$  and from the fixed parameters of the system  $A$ . Figures 5 and 6 show the corresponding results which have been computed with this method.

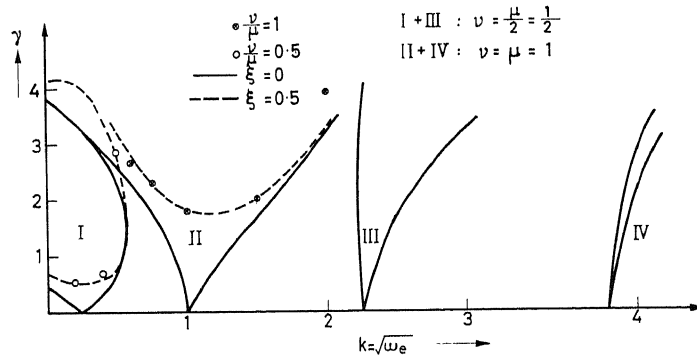


Figure 5. The amplitude of the parameter oscillation  $\gamma$  ( $\mu = \text{constant} = 1$ ) at the stability boundary, dependent on the natural frequency of a second-order system

In Figure 5 the value  $\gamma$  at the stability boundary has been plotted, dependent on the resonance frequency  $\omega_e = \sqrt{K}$  for zero damping and for  $\xi = 0.5$  with the value  $\mu = 1$  kept constant.  $\gamma$  is the amplitude of the parametric exciting oscillation. The meaning  $\omega_e = \sqrt{K}$  can be seen from eqn (3).

The plots for  $\xi = 0$  are known<sup>4</sup> as Ince-Strutt stability diagram. These diagrams can be computed easily with this method with the aid of the transfer function and the describing function method shown in this paper, with the consideration of the first and third harmonic for the value  $\nu/\mu = 0.5$  and with the consideration of the d.c. component, the first and the second harmonic for the ratio  $\nu/\mu = 1.0$ . The calculations for the damping factor  $\xi = 0.5$  are only slightly more complicated. It has been proved with the analogue computer that the influence of the higher harmonics is small and mostly negligible. The results found by analogue simulation are dotted in Figures 5

and 6. The deviation compared with the calculated values is negligible.\*

In Figure 6 the value  $\gamma$  dependent on the frequency of the varying parameter at the stability boundary has been plotted for a fixed parameter system  $A$  with the resonance frequency kept constant equal to one. The plot of Figure 6 demonstrates, that for  $\xi = 0.5$ , and for zero-damping, the measured and the calculated values are comparable for frequencies larger than  $\mu/\omega_e = 0.75$ . In Figure 7 the analogue computer plots  $\gamma(\tau)$  and  $\gamma(\tau)$  for  $\mu = 1.5$  ( $\nu = 0.5\mu$ ) and  $\mu = 0.75$  ( $\nu = \mu$ ) are shown.

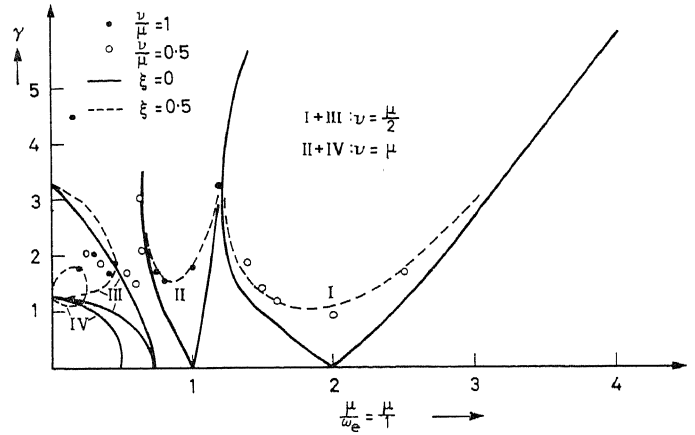


Figure 6. Amplitude  $\gamma$  of the parameter oscillation at the stability boundary of a second-order system with the natural frequency 1 as function of the frequency of the varying parameter

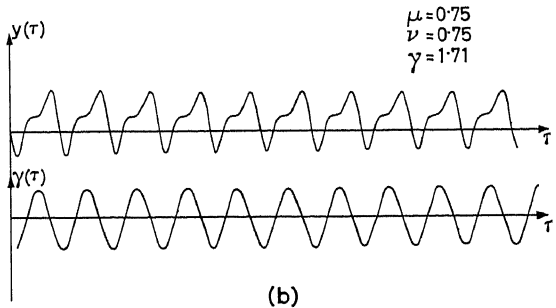
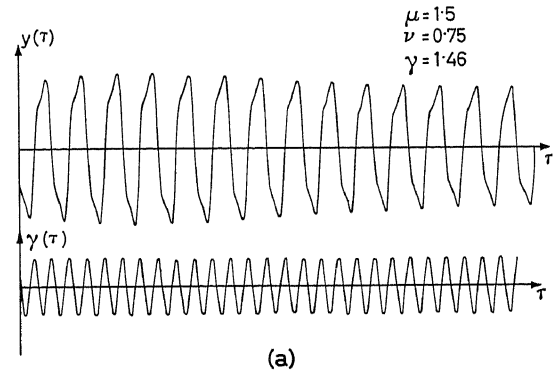


Figure 7. Undamped oscillations of a parametric excited system: (a)  $\mu = 1.5$ ,  $\nu = 0.5\mu$ ; (b)  $\mu = 0.75$ ,  $\nu = \mu$

\* Acknowledgement is due to Dr. Gerhard Schweizer, who performed all the analogue computations.

One finds that for  $\mu = 1.5$ , the third harmonic which has been considered in the calculation, and for  $\mu = 0.75$ , the d.c. component and the second harmonic which have been taken into account, are predominant. In Figure 8 the plots for  $\mu = 2$  ( $\nu = 0.5\mu$ ) and for  $\mu = 1$  ( $\nu = \mu$ ) are represented.  $y(\tau)$  is almost sinusoidal. A d.c. component will occur for  $\mu = 1$ .

The method cannot be applied successfully for the range of  $\mu$  from 0.1 to 0.6. Figure 6 does not give reliable comparisons

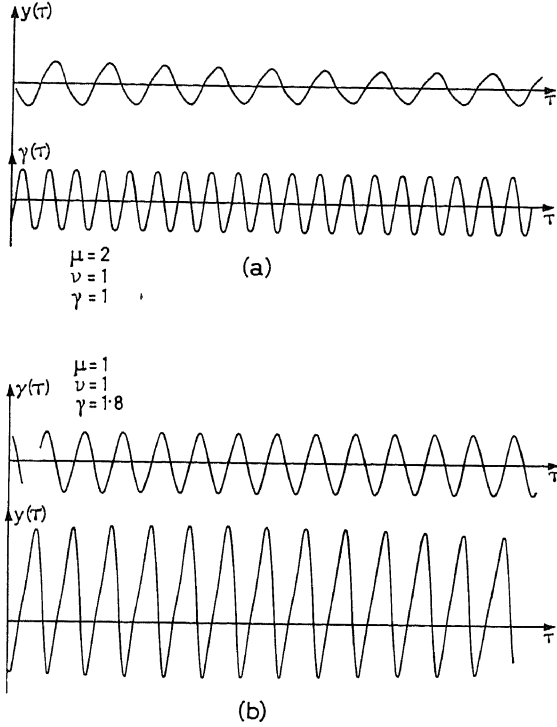


Figure 8. Undamped oscillations of a parametric excited system:  
(a)  $\mu = 2$ ,  $\nu = 0.5\mu$ ; (b)  $\mu = 1$ ,  $\nu = 1\mu$

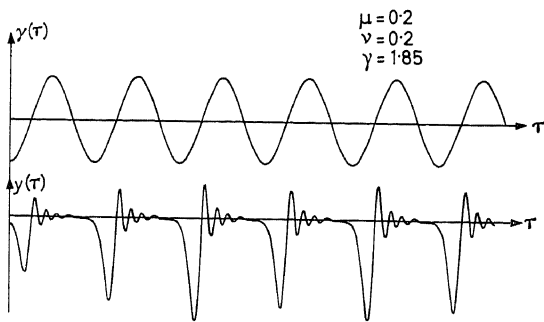


Figure 9. Undamped oscillations of a parametric excited system:  
 $\mu = 0.2$ ,  $\nu = 1\mu$

between the measured and calculated values. The reason is that many more harmonics should be considered. Figure 9 shows the plot for  $\gamma(\tau)$  and  $y(\tau)$  for the value  $\mu = 0.2$ . The parameter variation is so slow that the transients within one period will almost die out. During the next half period  $y(\tau)$  will rise about aperiodically. The final values of the second half period are the initial conditions for the third half period and so on. One

possible approach to the problem would be the replacement of the sinusoidal oscillation of the variable parameter by a square wave.

### Linear System with Two Periodically Varying Parameters

The analysis for time-varying systems has been limited up to now to the investigation of problems with one variable parameter. However, there are systems with two varying parameters. It is now shown with the aid of a specific problem—the oscillation of a swing—how this method may be applied.

It is well known that the amplitude of the oscillation of a swing is kept constant or excited by elevating or lowering the weight. This energy overcomes the unavoidable friction. Limiting oneself on the stability analysis, where the swing is oscillating with constant amplitude and the overall damping due to the friction of the bearings is kept zero by the energy of the variable parameters.

The oscillation of the swing can be represented by the differential equation where the length of the pendulum is varying<sup>5</sup>

$$\alpha'' mr^2 + \alpha'(\rho + 2rr'm) + mgr \sin \alpha = 0 \quad (29)$$

$m$  stands for the mass,  $r$  is the length of the pendulum,  $\alpha$  the angle of the pendulum,  $g$  is the gravity acceleration,  $\rho$  the damping (assumed to be proportional to the velocity).

The length of the pendulum  $r$  is varied by elevating and lowering the centre of gravity. A sinusoidal variation is assumed and one obtains

$$r = r_0 - \Delta r \sin(\omega_0 t + \psi) = r_0 [1 - \gamma \sin(\omega_0 t + \psi)] \quad (30)$$

$$r' = -r_0 \gamma \omega_0 \cos(\omega_0 t + \psi) \quad (31)$$

Eqn (29) can be written in the following form ( $\alpha$  is assumed to be small, which means  $\sin \alpha \approx \alpha$ , and  $T^2 = r_0/g$  is substituted)

$$\begin{aligned} \alpha'' T^2 + \alpha' \xi T + \alpha &= \alpha'' T^2 \gamma \sin(\omega_0 t + \psi) - \alpha' \xi T \gamma \sin(\omega_0 t + \psi) \\ &+ \alpha' T^2 2 \omega_0 \gamma \cos(\omega_0 t + \psi) \end{aligned} \quad (32)$$

$\xi$  is the normalized damping:

$$\xi = \frac{\rho}{mr_0 (gr_0)^{\frac{1}{2}}}$$

In the last equation, use was made of the following relation

$$\frac{1}{1 - \gamma \sin(\omega_0 t + \psi)} \approx 1 + \gamma \sin(\omega_0 t + \psi)$$

for small values of  $\gamma$ .

Introducing the normalized time  $\tau = t/T$  and  $\omega T = \Omega$ ,  $\omega_0 T = \Omega_0$  one obtains

$$\begin{aligned} \alpha'' + \alpha' \xi + \alpha &= \alpha'' \gamma \sin(\Omega_0 \tau + \psi) + \alpha' 2 \Omega_0 \gamma \cos(\Omega_0 \tau + \psi) \\ &- \alpha' \xi \gamma \sin(\Omega_0 \tau + \psi) \end{aligned} \quad (32a)$$

One knows from experience that one has to raise the weight at the lowest point of the swing oscillation, that means at  $\alpha = 0$ , if one intends to excite the swing oscillation. During one period of the oscillation one has therefore to raise and to lower its weight twice. The frequency of the parametric oscillation is therefore twice the frequency of the swing oscillation. The

easiest way, but not the only one, to keep the swing going, is to oscillate with the resonance frequency. Therefore, putting

$$\omega_0 = 2\omega_e \approx \frac{2}{T} \quad (33)$$

yields

$$\Omega_0 = \omega_0 T = 2 \quad (34)$$

According to eqn (32)  $1/T$  is the resonance frequency of the swing without consideration of the friction.  $\xi$  is always small compared with 1. The same is true for  $\gamma$ , therefore the variable  $\alpha' \xi \gamma \sin(\Omega_0 \tau + \psi)$  in eqn (32b) can be dropped.

At the boundary of the stability the amplitude of the oscillation is constant. Considering the fundamental harmonic and the third harmonic, one can write with  $\tau = t/T$ :

$$\left. \begin{aligned} \alpha_1 &= \hat{\alpha}_1 \sin \tau \\ \alpha_3 &= \hat{\alpha}_3 \sin(3\tau + \varphi) \end{aligned} \right\} \quad (35)$$

Then the following equations are obtained

$$\left. \begin{aligned} \alpha'_1 &= \hat{\alpha}_1 \cos \tau \\ \alpha'_3 &= 3 \hat{\alpha}_3 \cos(3\tau + \varphi) \end{aligned} \right\} \quad (36)$$

$$\left. \begin{aligned} \alpha''_1 &= -\hat{\alpha}_1 \sin \tau \\ \alpha''_3 &= -9 \hat{\alpha}_3 \sin(3\tau + \varphi) \end{aligned} \right\} \quad (37)$$

The aim now is to calculate the value  $\gamma = \Delta r/r_0$  so that an undamped oscillation can sustain.

Considering the fundamental harmonic and the third harmonic, one can perform the calculation with the aid of the block diagram of Figure 10. Part A of Figure 10 corresponds to the fixed parameter system, which is equivalent to the left side of eqn (32b). The blocks B and C are additional systems, which stand for the right side of eqn (32b), where both differentials are multiplied by the parametric oscillation. The feedback loop is now assumed to be open at the designated spot, but only for the fundamental harmonic. It is closed for the third harmonic.

According to Figure 10, the following relations can be obtained

$$\left. \begin{aligned} \beta_{B1} &= \alpha_1 B_{11}; \beta_{C1} = \alpha_1 C_{11}; \delta_{B1} = \alpha_3 B_{31}; \delta_{C1} = \alpha_3 C_{31} \\ \beta_{B3} &= \alpha_1 B_{13}; \beta_{C3} = \alpha_1 C_{13} \\ \alpha_1 &= \varepsilon_1 F_1; \alpha_3 = (\beta_{B3} + \beta_{C3}) F_3 \end{aligned} \right\} \quad (38)$$

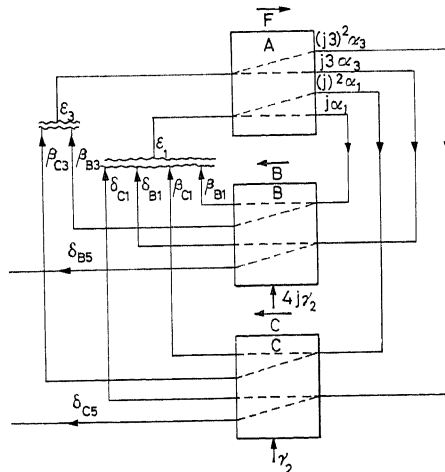


Figure 10. Block diagram for the determination of the stability boundary of a swing oscillation with two periodically varying parameters

For the stability boundary one gets

$$1 = \frac{\beta_{B1} + \beta_{C1} + \delta_{B1} + \delta_{C1}}{\varepsilon_1} \quad (39)$$

$$= F_1 [B_{11} + C_{11} + (B_{13} + C_{13}) F_3 (B_{31} + C_{31})]$$

$$= F_1 [B_{11} + C_{11} + F_3 (B_{13} B_{31} + C_{13} C_{31} + B_{13} C_{31} + C_{13} B_{31})]$$

Substituting  $\Omega = 0.5 \Omega_0 = 1$  the relations yield

$$F_1 = \frac{1}{(j\Omega)^2 + j\xi\Omega + 1} = \frac{1}{j\xi}$$

$$F_3 = \frac{1}{-9 + j3\xi + 1}$$

$$B_{11} = \frac{1}{2} \cdot 4\gamma (j \cos \psi - \sin \psi)$$

According to eqn (18) and eqn (9), considering  $2 \Omega_0 = 4$ , one obtains

$$C_{11} = -\frac{1}{2} \gamma (j \cos \psi - \sin \psi)$$

With the aid of eqns (9) and (32) the relations can be determined

$$B_{13} B_{31} = -12\gamma^2$$

$$C_{13} C_{31} = -\frac{9}{4} \gamma^2$$

$$C_{13} B_{31} = 3\gamma^2$$

$$B_{13} C_{31} = 9\gamma^2$$

Finally one will find

$$1 = \frac{1}{j\xi} \left[ \gamma (-1.5 \sin \psi + j 1.5 \cos \psi) - \gamma^2 \frac{2.25}{-8 + j3\xi} \right] \quad (40)$$

Neglecting for the moment the third harmonic, the part with  $\gamma^2$  drops. In this case one obtains the two equations  $\sin \psi = 0$  and  $\xi = 1.5 \gamma \cos \psi$ . With the value  $\xi = 0.1$  one calculates e.g.  $\gamma = 0.1/1.5 = 0.066$ . With  $\xi = 0.1$ , which corresponds to a damping time constant of the swing oscillation of

$$T_d = \frac{2T}{\xi} = 20T = \frac{20}{2\pi} T_p = 3T_p$$

the ratio of the amplitudes after three periods of the oscillation is 0.36. In this case  $\gamma$  must be  $\gamma = 0.07$ . From this it follows that the length of the pendulum has to be oscillated  $\pm 7$  per cent, so that an undamped oscillation can be sustained. If the third harmonic is considered, then the following two equations are obtained:  $\gamma \cdot 1.5 \sin \psi - \gamma^2 \cdot 0.28 = 0$  and  $0.1 - \gamma \cdot 1.5 \cos \psi - \gamma^2 \cdot 0.01 = 0$ .

Nearly the same value for  $\gamma$  is obtained from these two equations as before, without considering the third harmonic. It is assumed in Figure 10 that two parametric oscillations of the variable parameters are either in phase or  $90^\circ$  out of phase.

No difficulties arise if one allows an arbitrary phase angle between the two oscillations. In this case one has to introduce the fixed phase angle between the two parametric oscillations besides the phase angle  $\psi$  of the oscillation of the system and one of the parametric oscillations.

Even the two frequencies of the parametric oscillations must not be the same. But system oscillations will only occur, due to both variable parameters at certain ratios of the frequencies. The ratio must be such that each parameter oscillation, modulated by the system oscillation, yields one component of one frequency of the system oscillation. For a frequency ratio of 1:2 a system oscillation of the lower frequency can occur. *Figures 3 and 4* can be used as before for the analysis.

#### Arbitrary Linear System and Non-harmonic Parametric Oscillation

It can easily be seen from the stability analysis, so far, that investigations can be made for any linear system, which must not be of the second order, with periodically varying parameters. The linear part with fixed parameters is represented by the transfer function  $F(j\omega)$  in any case. The describing function remains the same.

If one parameter varies periodically but not harmonically, the stability analysis of each oscillation, determined by the Fourier series, can be investigated separately. In general the consideration of the fundamental harmonic will be sufficient.

#### References

- <sup>1</sup> LAUBER, R. Investigations into the stability of reactor control loops, including nonlinearities (in German). *Dissertation*. 1962. T. H. Stuttgart
- <sup>2</sup> LAUBER, R. A new procedure of the closed form calculation of describing functions of certain non-linear control loop elements. *Automatic and Remote Control*. 1963. London; Butterworths
- <sup>3</sup> SCHWEIZER, G. The closed form treatment of control loops with periodic parameters using operator calculus, *Automatic and Remote Control*, 1963, London; Butterworths
- <sup>4</sup> KLOTTER, H. Course in Technical Oscillations (in German). 1951. Berlin-Göttingen-Heidelberg; Springer
- <sup>5</sup> MAGNUS, K. Oscillations (in German). 1961. Stuttgart; B. G. Teubner

#### DISCUSSION

P. DORATO, *Polytechnic Institute of Brooklyn, 333, Jay Street, Brooklyn 1, N.J., U.S.A.*

I would like to point out some results obtained by my colleague Professor Bongiorno on the stability of periodically-varying linear systems presented at the 1962 I.A.C.C. Conference in New York. A typical result is as follows.

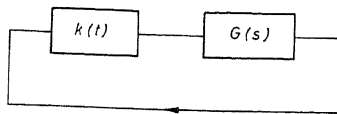


Figure A

The system shown in *Figure A* is stable for any periodically-varying gain  $k(t)$  if the condition

$$\max |G(j\omega)| \leq \frac{1}{k_{\max}}$$

where  $|k(t)| \leq k_{\max}$ . The structure of this figure is quite general since any time-varying parameter can, as shown, always be isolated as a gain. In addition, Professor Bongiorno has obtained estimates of the decay time when the system is stable.

A. LEONHARD, *in reply*

The case discussed by Professor Dorato—when the gain of the feedback loop  $k(t)$  is time-varying—is a special case, which can obviously be treated by the method exposed in my paper. The criterion cited by the Professor, which has been derived using methods differing from mine, appears, however, as a sufficient but not absolutely necessary condition for stability.

# On the Inverse Describing Function Problem

J. E. GIBSON and E. S. di TADA

## Summary

With the exception of a passing reference in an early note by Zadeh<sup>1</sup>, little or no attention appears to have been paid to the inverse describing function (IDF) problem. In essence this problem is concerned with extracting a description of a non-linear element given only its describing function. The problem is important as the final link in the non-linear synthesis problem for automatic control systems, for various experimental determinations, and finally for itself.

In this paper an analytic solution to the problem which results in expressions in the form of Volterra integral equations is provided. The conditions under which the representation for the non-linear element is unique are discussed and additional requirements for the unique definition of non-single-valued non-linear elements are given. As might be anticipated the describing function of a memory-type non-linear element is not sufficient to obtain a unique representation for the element.

Practical computational techniques are mentioned for the machine solution of the IDF and several examples are given.

## Sommaire

A part une allusion figurant dans un travail de Zadeh<sup>1</sup>, peu d'études ont été consacrées au problème de la transformation inverse de la fonction descriptive (inverse descriptive function (IDF)). Ce problème consiste à déterminer la fonction originale d'un élément non linéaire caractérisé par sa fonction descriptive; il se pose lorsqu'il s'agit de faire la synthèse de systèmes de réglage non linéaires dont les caractéristiques ont été déterminées expérimentalement; d'autre part, il présente un intérêt pour lui-même.

Ce rapport donne une solution analytique de ce problème résultant de l'application des équations intégrales de Volterra. Les conditions permettant de représenter un élément non linéaire sont indiquées et certains cas particuliers sont traités en détail avec indication des méthodes d'analyse à utiliser pour traiter ce problème au moyen de calculateurs numériques. Plusieurs exemples d'application sont donnés.

## Zusammenfassung

Mit Ausnahme einer beiläufigen Bemerkung in einer frühen Veröffentlichung von Zadeh<sup>1</sup> wurde bisher dem Problem der inversen Beschreibungsfunktion unseres Wissens wenig oder keine Aufmerksamkeit geschenkt. Im wesentlichen betrifft dieses Problem die Ableitung der Beschreibung eines nichtlinearen Gliedes, wenn nur dessen Beschreibungsfunktion angegeben ist. Das Problem ist (als letzter Schritt) bei der Synthese nichtlinearer Regelungen für die verschiedenen experimentellen Untersuchungen und auch für sich allein von Interesse.

Der Aufsatz enthält eine analytische Lösung des Problems, die auf Ausdrücke in Form von Volterraschen Integralgleichungen führt. Die Bedingungen, unter denen die Darstellung des nichtlinearen Gliedes eindeutig ist und die zusätzlichen Erfordernisse für eine eindeutige Definition eines nichtlinearen Gliedes mit mehrdeutiger Kennlinie werden angegeben. Wie zu erwarten, reicht die Beschreibungsfunktion eines nichtlinearen Gliedes mit Speicherwirkung nicht aus, um zu einer eindeutigen Darstellung zu gelangen.

Rechenverfahren zur maschinellen Lösung der inversen Beschreibungsfunktion werden erläutert und einige Beispiele angegeben.

## Introduction

The describing function technique has reached great popularity, principally because of the relative ease of computation involved

and the general usefulness of the method in engineering problems.

However, in the past, the describing function technique has been useful only in analysis. More exactly, it is a powerful tool for the investigation of the possible existence of limit cycles and their approximate amplitudes and frequencies.

Recently, synthesis techniques have begun to be developed for non-linear systems (Haussler<sup>2</sup>), in which the goal is to find the describing function of the element being synthesized. Therefore, for Haussler's method to be useful, a way must be found to reconstruct the non-linearity from its describing function. This is called the inverse describing function problem and is essentially a synthesis problem.

It is not the only case in which the IDF can be useful. Sometimes, in order to find the input-output characteristic of a physical non-linear element, a harmonic test can be easier to perform rather than a static one (which also may be insufficient). Finally, it is presumed that a solution to the IDF problem, as well as being interesting for its own sake, will add to the understanding of the conventional describing function.

In this paper an analytic solution to the IDF problem is presented and its uniqueness under various conditions discussed. A number of examples of practical calculations are included.

## Definitions

Consider a non-linear element and define the input and output variables as  $x$  and  $y$  respectively. The non-linear element will be defined only for a sinusoidal input

$$x = E \sin \omega t = E \sin \alpha \quad (1)$$

In general, the non-linear element might be allowed to be a double-valued function of  $x$ ,  $\dot{x}$  and  $E$

$$y = F(x, \dot{x}, E) \quad (2)$$

as, for example, would be necessary in a realistic, macroscopic model of magnetic saturation. In this paper the element will be considered to be independent of frequency, i.e.  $y = F(x, E)$ . If the output  $y$  satisfies the Dirichlet conditions, it can be expanded in a Fourier series

$$\begin{aligned} y(t) = & \frac{A_0(E)}{2} + A_1(E) \cos \alpha + A_2(E) \cos 2\alpha + \dots \\ & + A_n(E) \cos n\alpha + \dots + B_1(E) \sin \alpha + B_2(E) \sin 2\alpha + \dots \\ & + B_n(E) \sin n\alpha + \dots \end{aligned} \quad (3)$$

where  $A_n(E)$  and  $B_n(E)$  are the Fourier coefficients given by the following expressions

$$\begin{aligned} A_n(E) &= \frac{1}{\pi} \int_0^{2\pi} f(E \sin \alpha) \cos n\alpha \, d\alpha \\ B_n(E) &= \frac{1}{\pi} \int_0^{2\pi} f(E \sin \alpha) \sin n\alpha \, d\alpha \end{aligned} \quad (4)$$

The ratio of the  $n$ th harmonic amplitude to the amplitude of the input is defined as the  $n$ th describing function, a complex quantity in general,

$$K_{eq, n} = g_n(E) + j b_n(E) = \frac{B_n(E)}{E} + j \frac{A_n(E)}{E} \quad (5)$$

The conventional describing function results when  $n = 1$ , and the subscript is usually omitted.

### Integral Representation of the Describing Function

With  $y = F(x, E)$  as dealt with above, let  $F_1(x, E)$ , a single-valued though possibly discontinuous function, represent  $F(x, E)$  when  $\dot{x}$  is negative and  $F_2(x, E)$  when  $\dot{x}$  is positive. For later convenience let  $\alpha = \beta + \pi/2$ , then

$$\boxed{\text{Eqn (6)}} *$$

Let

$$\begin{aligned} F_1(x, E) &= P_1(x, E) + Q_1(x, E) \\ F_2(x, E) &= P_2(x, E) + Q_2(x, E) \end{aligned} \quad (7)$$

where  $P_1(x, E) = P_1(-x, E)$  and  $P_2(x, E) = P_2(-x, E)$  and  $Q_1(x, E) = -Q_1(-x, E)$  and  $Q_2(x, E) = -Q_2(-x, E)$ . Now the  $P$ 's and  $Q$ 's may be written in terms of either  $F_1$  or  $F_2$  with positive and negative arguments. The linear independence of the resulting sets of equations is easily checked. Now substitute (7) into (6), then let  $\beta = \gamma + \pi$  and after some elementary trigonometric transformations,

$$\boxed{\text{Eqn (8)}} \dagger$$

where

$$\begin{aligned} P(x, E) &= P_1(x, E) - P_2(x, E) \\ Q(x, E) &= Q_1(x, E) + Q_2(x, E) \end{aligned} \quad (9)$$

Since  $P_1$  and  $P_2$  are both even functions of  $x$ ,  $P$  is also an even function of  $x$ . For similar reasons  $Q$  is an odd function of  $x$ . Now let  $\theta = \pi - \gamma$  and employ the odd and even properties to obtain

$$\begin{aligned} g(E) &= \frac{2}{\pi E} \int_0^{\pi/2} Q(E \cos \theta, E) \cos \theta d\theta \\ b(E) &= \frac{-2}{\pi E} \int_0^{\pi/2} P(E \cos \theta, E) \sin \theta d\theta \end{aligned} \quad (10)$$

Now let  $E \cos \theta = x$ , then

$$\begin{aligned} g(E) &= \frac{2}{\pi E^2} \int_0^E \frac{x Q(x, E)}{(E^2 - x^2)^{1/2}} dx \\ b(E) &= -\frac{2}{\pi E^2} \int_0^E P(x, E) dx \end{aligned} \quad (11)$$

which are both in the form of Volterra integral equations.

The conventional method of computing the describing function of a non-linear element requires the knowledge of the actual shape of the output signal of the non-linear element when its input is driven by a sinusoidal wave. Then a Fourier analysis must be performed in order to find the amplitude of the first harmonic. This procedure is sometimes rather tedious, especially in the case in which the characteristic of the non-linear element is not known by an analytic expression but rather is given by experimental data. However, by using eqns (11) it is not necessary to compute the shape of the output, but only the two functions  $Q(x, E)$  and  $P(x, E)$ . These functions, given by eqns (9), can be computed directly from the characteristic of the non-linear element. This approach appears to possess an advantage over the original expression given by eqns (4) and (5). As a matter of fact, by means of eqns (11) a general method of machine computation of the describing function can be developed.

From the conceptual point of view, eqns (11) are interesting by themselves. With each single or double-valued non-linearity can be associated two single-valued functions which give the complete information about the non-linear element, in the sense that those two functions are sufficient to compute the describing function.

### Non-uniqueness of the Inverse Describing Function

#### Memory Type Non-linear Elements

To show the non-uniqueness of the solution of the integral eqn (11) (first half) for the case of memory-type non-linear

\* Eqn (6):

$$\begin{aligned} g(E) &= \frac{1}{\pi E} \left[ \int_0^\pi F_1(E \cos \beta, E) \cos \beta d\beta + \int_\pi^{2\pi} F_2(E \cos \beta, E) \cos \beta d\beta \right] \\ b(E) &= \frac{-1}{\pi E} \left[ \int_0^\pi F_1(E \cos \beta, E) \sin \beta d\beta + \int_\pi^{2\pi} F_2(E \cos \beta, E) \sin \beta d\beta \right] \end{aligned} \quad (6)$$

† Eqn (8):

$$\begin{aligned} g(E) &= \frac{1}{\pi E} \left[ \int_0^\pi P(E \cos \gamma, E) \cos \gamma d\gamma + \int_0^\pi Q(E \cos \gamma, E) \cos \gamma d\gamma \right] \\ b(E) &= -\frac{1}{\pi E} \left[ \int_0^\pi P(E \cos \gamma, E) \sin \gamma d\gamma + \int_0^\pi Q(E \cos \gamma, E) \sin \gamma d\gamma \right] \end{aligned} \quad (8)$$



elements, it is sufficient to show the existence of a set of functions  $Q_0(x, E)$ , not identically zero, whose corresponding  $g(E)$  are identically zero. Assume  $Q_0(x, E)$  to be of the form

$$Q_0(x, E) = h_1(x) m_1(E) + h_2(x) m_2(E) \quad (12)$$

and attempt to choose  $h_1(x)$ ,  $h_2(x)$ ,  $m_1(E)$  and  $m_2(E)$  in order to have  $g(E) = 0$ . Substituting eqn (12) into the first half of eqn (11)

$$m_1'(E) \int_0^E \frac{x h_1(x)}{(E^2 - x^2)^{\frac{1}{2}}} dx + m_2(E) \int_0^E \frac{x h_2(x)}{(E^2 - x^2)^{\frac{1}{2}}} dx = 0 \quad (13)$$

This means that if  $h_2(x)$ ,  $m_1(E)$  and  $m_2(E)$  are chosen arbitrarily (assuming that  $m_2(E)/m_1(E)$  has meaning)  $h(x)$  will be given by the solution of the following integral equation

$$\int_0^E \frac{x h_1(x)}{(E^2 - x^2)^{\frac{1}{2}}} dx = -\frac{m_2(E)}{m_1(E)} \int_0^E \frac{x h_2(x)}{(E^2 - x^2)^{\frac{1}{2}}} dx \quad (14)$$

Solving eqn (14) for  $h_1(x)$  (see Appendix) yields

$$h_1(x) = -\frac{2}{\pi x} \frac{d}{dx} \int_0^x dz \int_0^z \frac{z m_2(z) y h_2(y)}{m_1(z) (x^2 - z^2)^{\frac{1}{2}} (z^2 - y^2)^{\frac{1}{2}}} dy \quad (15)$$

Therefore to every function of the type \*

$$Q_0(x, E) = h_2(x) m_2(E) - \frac{2 m_1(E)}{\pi x} \frac{d}{dx} \int_0^x y h_2(y) \int_y^x \frac{z m_2(z) dz dy}{m_1(z) (x^2 - z^2)^{\frac{1}{2}} (z^2 - y^2)^{\frac{1}{2}}} \quad (16)$$

will correspond  $g(E) \equiv 0$ .

To illustrate the procedure one example should be considered. Let

$$\frac{m_2(E)}{m_1(E)} = E^2$$

Eqn (16) becomes

$$Q_0(x, E) \equiv h_2(x) m_2(E) - \frac{2 m_1(E)}{\pi x} \frac{d}{dx} \int_0^x y h_2(y) \int_y^x \frac{z^3}{(x^2 - z^2)^{\frac{1}{2}} (z^2 - y^2)^{\frac{1}{2}}} dz dy \quad (17)$$

but

$$\int_y^x \frac{z^3}{(x^2 - z^2)^{\frac{1}{2}} (z^2 - y^2)^{\frac{1}{2}}} dz = \frac{\pi}{4} (x^2 + y^2) \quad (18)$$

Substituting (18) into (17)

$$Q_0(x, E) = h_2(x) m_2(E) - \frac{m_1(E)}{2x} \frac{d}{dx} \int_0^x y h_2(y) (x^2 + y^2) dy \quad (19)$$

which can be reduced to

$$Q_0(x, E) = m_1(E) \left[ h_2(x) (E^2 - x^2) - \int_0^x y h_2(y) dy \right] \quad (20)$$

Suppose that  $m_1(E) = 1$  and  $h_2(x) = x$  then eqn (19) becomes

$$Q_0(x, E) = x E^2 - \frac{4}{3} x^3 \quad (21)$$

\* Because of the symmetry of eqn (13) the sub-indices 1 and 2 can be interchanged in eqn (16).

In Figure 1 is represented the block diagram of this non-linear element.

In an analogous manner the non-uniqueness of the solution of the first half of the integral eqn (11) can be demonstrated.

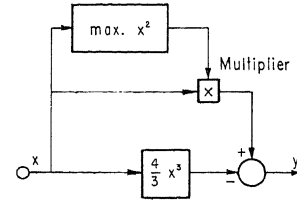


Figure 1. Example of a non-linear element with a describing function identically null

Non-memory Type (possibly double valued)

If the non-linear element is of the non-memory type, eqns (11) are reduced to

$$g(E) = \frac{2}{\pi E^2} \int_0^E \frac{x Q(x)}{(E^2 - x^2)^{\frac{1}{2}}} dx$$

$$b(E) = -\frac{2}{\pi E^2} \int_0^E P(x) dx \quad (22)$$

Both integral equations are Volterra integral equations of the first kind and their solutions will be unique†. Therefore, given the functions  $g(E)$  and  $b(E)$  there will exist one and only one pair of functions  $Q(x)$  and  $P(x)$  that, substituted in eqn (22), will transform these equations in an identity. However  $P(x)$  and  $Q(x)$  are not sufficient to determine the non-linearity. As a matter of fact, the equation of the non-linearity will only be determined if  $P_1(x)$ ,  $P_2(x)$ ,  $Q_1(x)$  and  $Q_2(x)$  are known. From eqn (9) it can be shown that, given  $P(x)$  and  $Q(x)$ , any set of equations  $P_1(x)$ ,  $P_2(x)$ ,  $Q_1(x)$  and  $Q_2(x)$  that satisfy eqn (9), will generate a different non-linearity with the same describing function. Therefore, the knowledge of the describing function of a non-linear element is not sufficient to determine the equation of the non-linear element. Even in the case of single-valued non-linear elements,  $g(E)$  is not sufficient to determine the non-linearity. However, if in addition the even harmonics of the Fourier series representation are zeros, the non-linear element is symmetric and  $g(E)$  is sufficient.

### Inverse Describing Function

The solution of the pair of integral equations which represents a non-linear element of the non-memory type is not far to seek. In the first half of eqn (22) let  $v = x^2$  and  $\eta = E^2$ ,

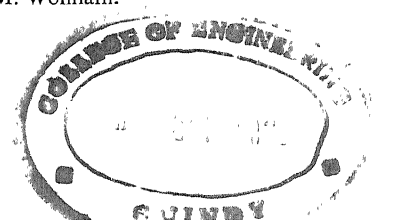
$$\pi \eta g(\sqrt{\eta}) = \int_0^\eta \frac{Q(\sqrt{v})}{(\eta - v)^{\frac{1}{2}}} dv \quad (23)$$

which is Abel's equation‡. The solution of (23) is well known<sup>3</sup>, and may be written in terms of the original variables as

$$Q(x) = \int_0^{x^2} \frac{d[z^2 g(z)]}{(x^2 - z^2)^{\frac{1}{2}}} \quad (24)$$

†  $Q(x)$  and  $P(x)$  are assumed to be continuous functions.

‡ Pointed out to authors by W. M. Wonham.



The solution of the second half of eqn (22) may be written directly, provided  $d/dx b(x)$  exists, as

$$P(x) = -\frac{\pi}{2} \frac{d}{dx} [x^2 b(x)] \quad (25)$$

### Examples

(1) Consider the following describing function

$$g(E) = \frac{3}{4} E^2; \quad b(E) = 0 \quad (26)$$

A more convenient form perhaps for calculation of eqn (24) is

$$Q(x) = \frac{d}{x dx} \int_0^x \frac{z^3 g(z)}{(x^2 - z^2)^{\frac{3}{2}}} dz \quad (27)$$

which for this example becomes

$$\boxed{\text{Eqn (28)}}^*$$

and if  $F(x)$  is assumed to be single-valued and symmetric, then

$$\frac{1}{2} Q(x) = Q_1(x) = Q_2(x) = f_1(x) = f_2(x) = f(x).$$

Therefore

$$f(x) = x^3 \quad (29)$$

(2) Consider the following describing function

$$\begin{aligned} g(E) &= 0 & (E < a) \\ g(E) &= \frac{4M}{\pi E^2} (E^2 - a^2)^{\frac{1}{2}} & (E \geq a) \end{aligned} \quad (30)$$

$$b(E) = 0$$

In order to apply eqn (24), write

$$\frac{d}{dz} [z^2 g(z)] = \frac{4M}{\pi} \frac{z}{(z^2 - a^2)^{\frac{1}{2}}} \quad (31)$$

Therefore

$$Q(x) = \frac{4M}{\pi} \int_a^x \frac{z dz}{(z^2 - a^2)^{\frac{1}{2}} (x^2 - z^2)^{\frac{3}{2}}} \quad (32)$$

and

$$\begin{aligned} Q(x) &= 2M & (x \geq a) \\ Q(x) &= 0 & (x < a) \end{aligned} \quad (33)$$

From eqn (25),

$$P(x) = 0 \quad (34)$$

If the non-linearity is assumed to be symmetric and single-valued it may be shown that,

$$f(x) = \frac{Q(x)}{2} = Mu(|x| - a) \operatorname{sgn} x \quad (35)$$

where  $u(x)$  is the step function.

In Figure 2, eqn (35) is plotted. As could have been foreseen,

the non-linear element which has the describing function given in eqn (30) is a relay with dead band.

Table 1. Inverse Describing Function of the Describing Function Shown in Figures 3 and 4  
(Experimental and theoretical results)

$x$	$F_1(x)$		$F_2(x)$	
	Experimental	Theoretical	Experimental	Theoretical
-5.00	-4.0017	-4.00	-3.9972	-4.00
-4.80	-3.9920	-4.00	-4.0045	-4.00
-4.60	-3.9971	-4.00	-3.9993	-4.00
-4.40	-4.0047	-4.00	-3.9921	-4.00
-4.20	-3.9951	-4.00	-3.9954	-4.00
-4.00	-3.9912	-4.00	-4.0127	-4.00
-3.80	-3.6030	-3.60	-3.6052	-3.60
-3.60	-3.2020	-3.20	-3.2079	-3.20
-3.40	-2.8027	-2.80	-2.9505	-2.95
-3.20	-2.4020	-2.40	-2.8516	-2.85
-3.00	-2.0034	-2.00	-2.7500	-2.75
-2.80	-1.6020	-1.60	-2.6537	-2.65
-2.60	-1.2025	-1.20	-2.5528	-2.55
-2.40	-0.80520	-0.80	-2.4521	-2.45
-2.20	-0.40372	-0.40	-2.3542	-2.35
-2.00	-0.00499	0.00	-2.2475	-2.25
-1.80	-0.00130	0.00	-2.1518	-2.15
-1.60	-0.00100	0.00	-2.0517	-2.05
-1.40	-0.02379	0.00	-1.8228	-1.80
-1.20	-0.00046	0.00	-1.4007	-1.40
-1.00	-0.00082	0.00	-1.0008	-1.00
-0.800	-0.00140	0.00	-0.60107	-0.60
-0.600	-0.00202	0.00	-0.20202	-0.20
-0.400	0.00000	0.00	0.00000	0.00
-0.200	0.00000	0.00	0.00000	0.00
0.00	0.00000	0.00	0.00000	0.00
0.20	0.00000	0.00	0.00000	0.00
0.40	0.00000	0.00	0.00000	0.00
0.60	0.20202	0.20	0.00202	0.00
0.80	0.60107	0.60	0.00141	0.00
1.00	1.0008	1.00	0.00082	0.00
1.20	1.4007	1.40	0.00046	0.00
1.40	1.8228	1.80	0.02379	0.00
1.60	2.0517	2.05	0.00100	0.00
1.80	2.1518	2.15	0.00130	0.00
2.00	2.2475	2.25	0.00499	0.00
2.20	2.3542	2.35	0.40372	0.40
2.40	2.4521	2.45	0.80520	0.80
2.60	2.5528	2.55	1.2025	1.20
2.80	2.6537	2.65	1.6020	1.60
3.00	2.7500	2.75	2.0034	2.00
3.20	2.8516	2.85	2.4020	2.40
3.40	2.9505	2.95	2.8027	2.80
3.60	3.2079	3.20	3.2020	3.20
3.80	3.6052	3.60	3.6030	3.60
4.00	4.0127	4.00	3.9912	4.00
4.20	3.9954	4.00	3.9951	4.00
4.40	3.9921	4.00	4.0047	4.00
4.60	3.9993	4.00	3.9971	4.00
4.80	4.0045	4.00	3.9920	4.00
5.00	3.9972	4.00	4.0017	4.00

\* Eqn (28):

$$Q(x) = \frac{1}{x} dx \left\{ \frac{3}{4} \left[ -\frac{(x^2 - z^2)^{\frac{5}{2}}}{5} + \frac{2(x^2 - z^2)^{\frac{3}{2}}}{3} - x^4 (x^2 - z^2)^{\frac{1}{2}} \right] \right\}_0^x = 2x^3 \quad (28)$$

(3) A numerical method to perform the operations indicated by eqns (24) and (25) has been proposed<sup>4</sup>. The philosophy of this method consists in dividing the interval in which the describing function is known in sub-intervals and approximating  $g(E)$  and  $b(E)$  by a polynomial in each one of these sub-intervals.

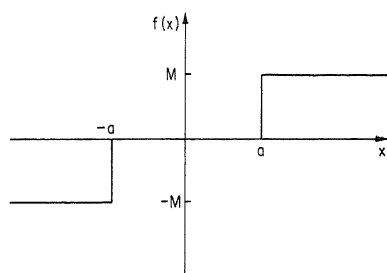


Figure 2. Plot of eqn (35)

In the work by di Tada<sup>4</sup> a second degree polynomial was used for the differentiations, and a linear approximation for the integrations.

In Figures 3, 4 and 5 one example is shown. In Table 1 the experimental and theoretical results are listed.

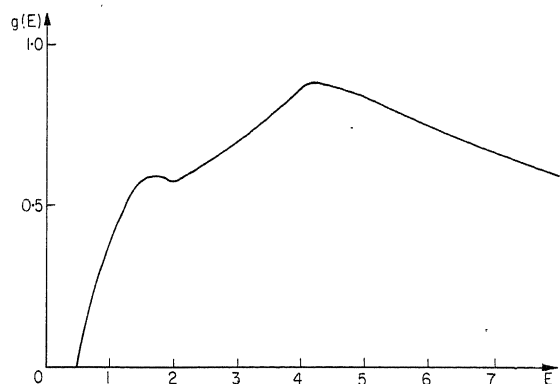


Figure 3. Real portion of the describing function of an unknown non-linear element

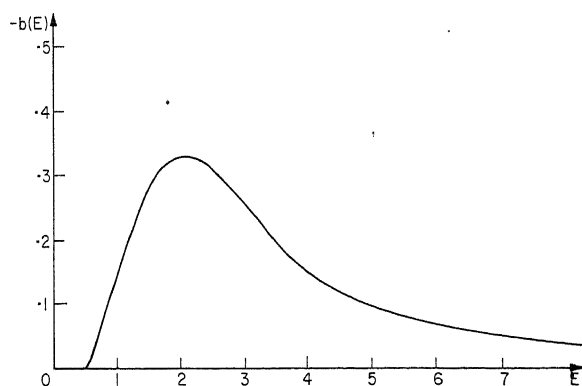


Figure 4. Imaginary portion of the describing function of an unknown non-linear element

## Conclusions

An analytical approach to the inverse describing function problem has been proposed. The problem is formulated in terms of certain integral equations of the Volterra type. The required

solutions are given. However, the operations expressed in eqns (24) and (25) are in general very difficult to perform in practical cases, if not impossible. In addition to this analytical difficulty, it should be pointed out that in the majority of practical cases the describing function is not given by an analytical ex-

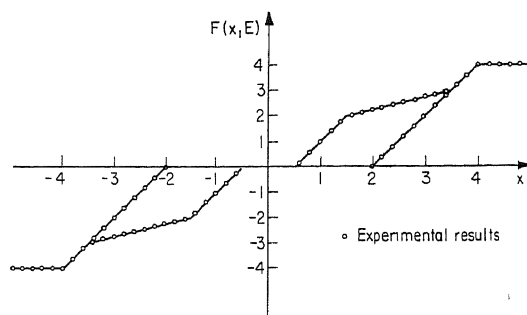


Figure 5. Calculated results for the non-linear element whose describing function is given by Figures 3 and 4

pression, but rather is given by experimental data. For these reasons it is suggested that numerical computation procedures be used in practice. Such numerical procedures are presently available<sup>5</sup>.

## References

- 1 ZADEH, L. A. On the identification problem. *Trans. Inst. Radio Engrs.* Vol. CT-3 (1956) 277-281
- 2 HAUSLER, R. L. A graphical method for finding the closed-loop frequency response of system with non-single valued nonlinearities. *CISL Memo.*, Purdue University, Indiana
- 3 WHITTAKER, E. T. and WATSON, G. N. *Modern Analysis* 4 ed. p. 229, 1950. London; Cambridge University Press
- 4 DI TADA, E. S. Analytical approach to the inverse-describing function problem. *M.S. Thesis* June 1962. Faculty of Electrical Engineering, Purdue University, Indiana
- 5 GIBSON, J. E., et al. Describing function inversion: theory and computational techniques. *Tech. Rep. EE62-10*, Control and Inf. Syst. Lab. Purdue University, Indiana, USA

## Appendix

The solution of Abel's integral equation

$$F(x) = \int_0^x \frac{f(z)}{(x^2 - z^2)^\alpha} dz \quad 0 < \alpha < 1 \quad (1)$$

may be written as

$$f(\xi) = \frac{2 \sin \alpha \pi}{\pi} \frac{d}{d\xi} \int_0^\xi \frac{x F(x)}{(\xi^2 - x^2)^{1-\alpha}} dx \quad (2)$$

Therefore for

$$\alpha = \frac{1}{2}$$

$$f(z) = z h_1(z)$$

$$F(x) = -\frac{m_2(x)}{m_1(x)} \int_0^x \frac{\eta h_2(\eta)}{(x^2 - \eta^2)^{\frac{1}{2}}} d\eta \quad (3)$$

Eqn (2) becomes

$$h_1(x) = -\frac{d}{dx} \frac{d}{dx} \int_0^x \frac{z m_2(z) \eta h_2(\eta) d\eta}{\pi x \int_0^z \frac{z m_2(z) \eta h_2(\eta) d\eta}{m_1(z) (z^2 - \eta^2)^{\frac{1}{2}} (x^2 - z^2)^{\frac{1}{2}}}}$$

## DISCUSSION

## Author's Opening Remarks

Between the submission of this paper and the I.F.A.C. Congress we have prepared standard FORTRAN programmes for the describing function and inverse describing function computations based on the principles given in the paper, and these routines are in regular use in our laboratory.

L. R. YOUNG, *Massachusetts Institute of Technology, 77, Massachusetts Avenue, Cambridge 39, Mass., U.S.A.*

The authors are to be congratulated for their presentation of a new and computationally simpler method of describing function determination, and exposition of the equations for determining the form of a non-linearity by means of the inverse describing function. The latter technique should prove valuable in experimental identification of non-linear control elements, especially in biological control systems. Biological adaptation rarely permits the use of static test inputs, but sinusoidal signals and harmonic analysis may be useful in exploring the non-linearity through its inverse describing function.

Some questions arise concerning the limitations of the method discussed in the paper. First, the exclusion of all but gain type non-linearities from consideration eliminates the important class of problems in which the frequency-dependent dynamics cannot be separated from the non-linearity in obtaining experimental data. Was this exclusion made for mathematical expediency or because of a fundamental limitation in the use of the inverse describing function with frequency-dependent non-linearities?

Second, does the existence of a memory type non-linearity with non-unique describing function imply that no useful information (in terms of an inverse describing function) can be derived for *any* memory type non-linearity?

Finally, is it always necessary to make some assumption about the symmetry or type of asymmetry of the non-linearity in order to determine a unique equation of the element from its describing function?

J. E. GIBSON, *in reply*

Dr. Young's remarks on the usefulness of the inverse describing function in the analysis of biological systems are appreciated. Let us make clear the classes of non-linear elements to which our approach applies. For computation convenience, we exclude those elements which are dependent on frequency. By memory-type elements we mean those in which the non-linear characteristic is a function of the amplitude of the driving signal. An example which comes to mind is the hysteresis loop of magnetic material. Here the hysteresis loop size depends on the amplitude of the driving signal. By non-memory elements we mean those with a fixed characteristic, for example as shown in Figure 5 of the paper, which may be multi-valued but which do not depend on the amplitude of the driving signal. This is not usually referred to as a gain-type non-linear element.

The fact that the inverse describing function in certain cases is not unique certainly does not imply that it is useless in those cases. As

shown in Reference 5 of the paper, one additional even harmonic describing function term is usually all that is required to make a unique identification of the non-linear element. This same approach applies if the element is not symmetrical.

YA. Z. TSYPKIN, *Institute of Automatic and Telemechanics, Kalanchevskaya 15a, Moscow, 1-15, U.S.S.R.*

The paper shows that the accurate determination of the characteristics of non-linear element by means of describing functions is dependent on the solution of the integral equation, which, in most cases, is somewhat slow. This task may be quite easily solved on the basis of the approximate expression of describing functions as given in References 1 and 2. For the single-valued characteristics of the non-linear element, we get

$$g(A) \approx \frac{2}{3A} \left[ F_1(A) - F_1\left(\frac{A}{2}\right) \right] \quad (1)$$

Substitute under (1)  $A$  for  $A/2^k$  ( $k = 0, 1, 2, \dots$ ) and multiply it by  $(-1)^k$ , there will be a series of equations. Starting from the addition of these equations, we obtain

$$F_1(A) = \sum_{k=0}^{\infty} (-1)^k \frac{3A}{2^{k-1}} g\left(\frac{A}{2^k}\right)$$

This formula is very easy to calculate owing to the rapid convergence of the series.

From the table<sup>1, 2</sup> it is, in the same way, possible to get the expression  $F(A)$  for more difficult cases.

## References

- <sup>1</sup> TSYPKIN, YA. Z. On the relation between the equivalent gain constant of the non-linear element and its characteristics. *Automatika i Telemekhanika* 17, No. 6 (1956)
- <sup>2</sup> TSYPKIN, YA. Z. On the relation between the characteristics of a non-linear element and its describing function. *Regelungstechnik* No. 8 (1958)

J. E. GIBSON and E. I. DI TADA, *in reply*

We thank Professor Tsyppkin for mentioning his early work on the approximate solution of eqn (11) of the paper. This work was brought to our attention only a few weeks ago by R. A. Johnson, and definitely deserves to be cited as a reference. We should also cite Mathews<sup>1</sup> for his approximate solution of the form given in eqn (10) of the paper. As we remark in the text, this is a convenient method of calculating describing functions. In this paper, however, we concentrate our attention on the inverse problem.

## Reference

- <sup>1</sup> MATHEWS, M. V. A method for evaluating non-linear servo mechanisms. *Applications and Industry*. (May 1955) 117-123

# Relative Stability of Oscillations in Non-linear Control Systems

Z. BONENN

## Summary

The relative stability of oscillations in non-linear systems is investigated using the incremental input describing function. The results can be used to obtain the response to small step inputs, and the method is illustrated by an example of an 'oscillating servomechanism'.

General formulae are given for calculating the incremental input describing function.

## Sommaire

La stabilité relative des oscillations dans des systèmes non linéaires est examinée à l'aide de la fonction descriptive à signal d'entrée échelonné. Les résultats peuvent être utilisés afin d'obtenir la réponse aux fonctions l'échelon de petites amplitudes. La méthode est illustrée par exemple d'un «servomechanisme oscillant».

Des formules générales, permettant de calculer la fonction descriptive à signal d'entrée échelonné, sont données.

## Zusammenfassung

Die relative Stabilität von Schwingungen in nichtlinearen Systemen wird mittels der „inkrementalen“ Beschreibungsfunktion untersucht. Die Ergebnisse können dazu dienen, die Antwort auf kleine Sprunganregungen zu bestimmen. Als Beispiel zur Erläuterung dieses Verfahrens dient ein „schwingender Regelkreis“.

Allgemeine Formeln zur Berechnung der inkrementalen Beschreibungsfunktion werden angegeben.

## Introduction

The stability of oscillations in non-linear systems may be investigated in various ways. One method, suggested by West<sup>1, 2</sup>, is to assess the stability according to the Nyquist criterion by plotting the open-loop frequency response locus for an incremental signal. The application of this method requires prior calculation of the incremental input describing function which may be calculated in a straightforward manner from the usual describing function<sup>3, 4</sup>.

In this paper, it is shown that by application of the describing function for an incremental input it is possible to obtain the stability and also the relative stability of oscillations. This solution can be used to obtain the response to small step inputs which may be considered as a perturbation to the steady-state oscillation.

## Theory

The standard mathematical approach to the stability problem of periodic oscillations in a non-linear system may be summarized as follows<sup>5</sup>. Suppose the system is represented by the differential equation (Figure 1)

$$x + G(D) \cdot f(x) = 0 \quad (1)$$

where  $G(D)$  is a linear differential operator with  $D = d/dt$  and  $f(x)$  is a non-linearity. When eqn (1) has a periodic steady-state

solution  $x_{ss}$  at frequency  $\omega_{ss}$ , the stability of this solution may be investigated by forming the first variation of eqn (1) with respect to  $x_{ss}$ . This variation is given by

$$\Delta x + G(D) \left( \frac{df}{dx} \right)_{x=x_{ss}} \cdot \Delta x = 0 \quad (2)$$

$\Delta x$  is the perturbation, the behaviour of which will determine the stability of  $x_{ss}$ .

Eqn (2) is a linear differential equation with periodic coefficients, the solution of which may be reduced to that of a set of linear algebraic equations. This set, however, is usually an infinite set and its solution to a sufficient degree of accuracy is usually a lengthy and difficult matter.

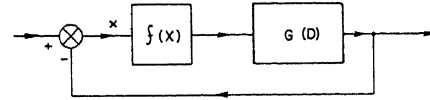


Figure 1. Non-linear system

For on-off control systems the above approach has been reduced to a neat form by Tsytkin<sup>6</sup>, whereby the variational equation of the on-off system is that of a linear sampled data system, the stability of which may be determined by known techniques. This method becomes quite complicated when the periodic oscillations are not simple, furthermore it does not apply to continuous non-linearities.

It appears that one of the main reasons for the practical difficulties in applying the above mathematical approach is that no restriction is placed beforehand on the perturbing signal, which may take any form whatsoever.

The solution of eqn (2) may be simplified by assuming a specified form of incremental perturbing signal. This method is analogous to the solution of eqn (1) by the describing function method (where by assuming a simple form for the steady-state solution, namely

$$x_{ss} = a e^{j\omega_{ss} t} \quad (3)$$

and neglecting components of  $x_{ss}$  at other frequencies, the solution is greatly simplified).

In the same manner, the solution of eqn (2) may be simplified by testing the stability of the steady-state oscillation with respect to a specified incremental perturbing signal

$$\Delta x = \Delta a \cdot e^{\sigma t} \cdot e^{j(\omega t + \gamma)} \quad (4)$$

where  $\Delta a = \text{constant}$ ,  $\omega$  varies from 0 to  $\infty$  and  $\gamma$  is an arbitrary phase angle. Using the theorem<sup>7</sup>

$$G(D) e^{\sigma t} f(t) = e^{\sigma t} G(D + \sigma) \cdot f(t) \quad (5)$$

eqn (2) may be written as

$$\Delta a e^{j(\omega t + \gamma)} + G(D + \sigma) \left( \frac{df}{dx} \right)_{x=x_{ss}} \cdot \Delta a e^{j(\omega t + \gamma)} = 0 \quad (6)$$

The first term is at frequency  $\omega$ . In the second term,

$$\left(\frac{df}{dx}\right)_{x=x_{ss}} \cdot \Delta a \cdot e^{j(\omega t + \gamma)}$$

is the output of the non-linearity due to the incremental signal and contains components at  $\omega$  and at other frequencies. Here the usual describing function approximation is made by neglecting output components at other frequencies and replacing the output term by

$$\Delta a N_i(a) e^{j(\omega t + \gamma)} \quad (7)$$

Here  $N_i(a)$  is the incremental input describing function which is defined by

$$N_i(a) = \frac{\text{output component at } \omega \text{ due to } \Delta a}{\Delta a} \quad (8)$$

$N_i(a)$  is calculated for  $a$  (i.e. for the steady-state amplitude, the stability of which is under investigation). Note that  $\Delta a$  and  $N_i(a)$  are constants. Hence, using eqns (6), (7) and (5) the following is obtained.

$$1 + G(\sigma + j\omega) N_i(a) = 0 \quad (9)$$

Eqn (9) is an extension of the equation used by West to determine the stability of oscillations.

$$1 + G(j\omega) \cdot N_i(a) = 0 \quad (10)$$

This extension is valid because eqn (10) is a linear equation [ $N_i(a)$  is a constant for the given ( $a$ )], hence  $G(j\omega)$  can be replaced by  $G(\sigma + j\omega)$ . From the block diagram (Figure 2) corre-

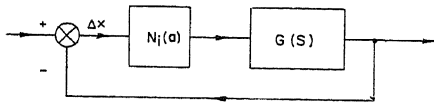


Figure 2. Incremental system

sponding to eqn (9) it is seen that the effect of the non-linearity on the perturbing signal is expressed by the equivalent gain  $N_i(a)$ . Also, it is obvious that the transition from eqn (10) to eqn (9) is exactly analogous to the transition from stability determination by the Nyquist criterion to the calculation of the closed-loop poles by the root-locus method.

Eqn (9) yields  $n$  solutions for the closed-loop poles,  $\sigma_1 + j\omega_1, \dots, \sigma_n + j\omega_n$  [ $n$  is the order of eqn (1)]. The complete solution for the perturbation is given by a combination of the  $n$  elementary solutions whose relative amplitudes are determined by the initial conditions. The damping factors ( $\sigma$ ) indicate the rate of decay or otherwise of the initial perturbation and thus determine the relative stability of the steady-state oscillation. Also, the solutions of eqn (9) can be used to find the response to small step inputs which may be considered as a perturbation to the steady-state oscillation.

Among the  $n$  solutions of eqn (9) there are at least two solutions at the steady-state frequency  $\omega = \omega_s$ . One of these solutions is with  $\sigma = 0$ . This solution, which does not decay, corresponds to a shift on the limit cycle and demonstrates the fact that the limit-cycle oscillation possesses only orbital stability but not asymptotic orbital stability<sup>5</sup>.

As for the other solutions, due to the filtering assumption used in describing function analysis, only solutions with  $\omega < \omega_s$  are considered. These solutions are always non-synchronous in odd non-linearities<sup>8</sup>.

The practical application of this method depends on the ready availability of  $N_i$ . Simple general formulae which express  $N_i$  directly in terms of the ordinary describing function are given in the final section. Both forms of  $N_i$ , synchronous and non-synchronous, are required for the complete solution of eqn (9).

### Example—Oscillating Servomechanism

Consider the system shown in Figure 3. Using conventional describing function analysis it is easily shown that the system has a stable steady-state oscillation  $E = a \sin \omega_s t$  where  $\omega_s = 10\frac{1}{2}$  (Figure 4). MacColl has shown<sup>9</sup> that these oscillations effectively linearize the relay for small, low frequency (compared with  $\omega_s$ ) input signals. Hence, he designated such a system as an 'oscillating servomechanism'.

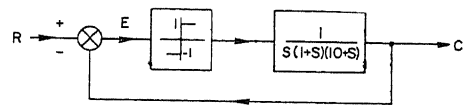


Figure 3. Oscillating servomechanism

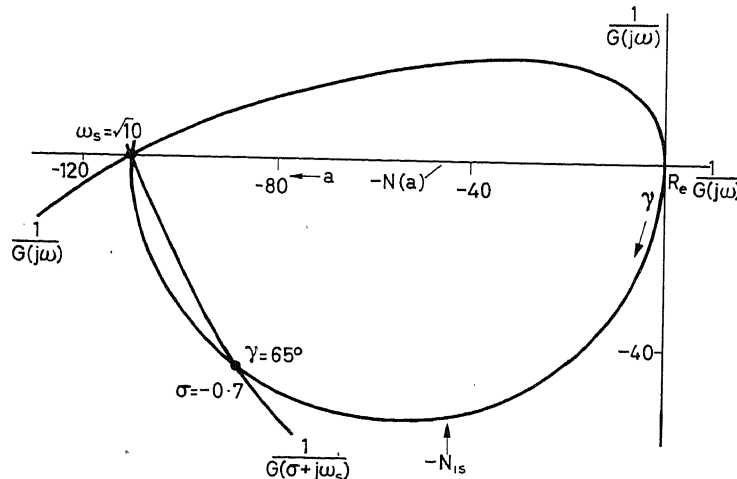


Figure 4. Solution for synchronous roots

The response of this system to a small step input may be obtained by viewing this input as a perturbation of the steady-state oscillation which decays according to the roots of eqn (9). First obtain the solution in the synchronous case where eqn (9) becomes

$$1 + N_{is}G(\sigma + j\omega_s) = 0 \quad (11)$$

Using  $N_{is}$  for the ideal relay [eqn (25)] two synchronous solutions are obtained (Figure 4)

$$s_1 = j(10^{\frac{1}{2}}), \gamma_1 = 90^\circ; \quad s_2 = -0.7 + j(10^{\frac{1}{2}}), \gamma_2 = 65^\circ \quad (12)$$

The first solution corresponds to a phase shift of the steady-state oscillation which does not decay. The second solution has an angle  $\gamma = 65^\circ$  with respect to the steady-state oscillation. The third solution is given by

$$1 + N_{ins}G(\sigma + j\omega_s) = 0 \quad (13)$$

This is a non-synchronous solution. Note that for the ideal relay [eqn (25)]

$$N_{ins} = \frac{N}{2} = -\frac{1}{2G(j\omega_s)} \quad (14)$$

Hence [eqns (13), (14)]

$$G(\sigma + j\omega_s) = 2|G(j\omega_s)| < -180^\circ \quad (15)$$

This equation is easily solved by root locus techniques yielding

$$s_3 = -10.55 \quad (16)$$

The complete solution for the perturbation is given by [eqns (12), (16)]

$$\Delta E(t) = \Delta a_1 \cos 10^{\frac{1}{2}}t + \Delta a_2 e^{-0.7t} \sin(10^{\frac{1}{2}}t + 65^\circ) + \Delta a_3 e^{-10.55t} \quad (17)$$

Assume, for example, a small step input  $\Delta R$  at  $t = 0$ . Using the initial conditions

$$\Delta E(0) = \Delta R; \quad \Delta E'(0) = \Delta E''(0) = 0 \quad (18)$$

$\Delta E(t)$  is given by

$$\Delta E(t) = \Delta R [-0.46 \cos 10^{\frac{1}{2}}t + 1.5 e^{-0.7t} \sin(10^{\frac{1}{2}}t + 65^\circ) + 0.1 e^{-10.55t}] \quad (19)$$

The response is dominated by the first and second term, which represent the phase and amplitude perturbations of the steady-state oscillation.

### Relation to Previous Work

Two other approaches to the problem of the relative stability of the steady state will be outlined. Starting from the conventional describing function solution  $1 + N(a)G(j\omega_s) = 0$ , Grensted<sup>10, 11</sup> assumed that the perturbed oscillation is at a variable complex frequency near the steady-state frequency

$$-\mu + j\omega = j\omega_s + \Delta z \quad (20)$$

From this assumption he derived a rather complicated variational equation which he solved for the perturbation modes. Two solutions equivalent to our synchronous solutions are obtained. However, due to the above assumption [eqn (20)] the non-synchronous solutions cannot be obtained.

Cosgriff's approach<sup>12</sup> is quite similar to the author's, assuming the same type of incremental signal [eqn (4)] to be used in the solution of eqn (2). However, he does not use the concept of the incremental input describing function. This concept enables the author to formalize the solution [eqn (9)] and clarify it. Also, as  $N_i$  can be calculated very easily from the ordinary describing function the solution becomes a routine matter.

### The Incremental Input Describing Function for the Ideal Relay

#### Synchronous Case

The input to the non-linearity  $f(x)$  is assumed to be

$$x = a e^{iq} + \Delta a e^{i(q+\gamma)} \quad (21)$$

where  $a$  is the steady-state amplitude and  $\Delta a \ll a$  is the perturbation.  $N_{is}$ , the synchronous incremental input describing function, is given by<sup>3</sup>

$$N_{is} = N + \frac{a}{2} \cdot \frac{dN}{da} (1 + e^{-2j\gamma}) \quad (22)$$

where  $N$  is the ordinary describing function. When  $N_{is}$  is plotted in the complex plane as a function of  $\gamma$  with  $a$  as parameter a family of circles is obtained. When  $\gamma = 90^\circ$ ,  $N_{is} = N$ .

The solution of eqn (9) in the synchronous case yields at least two intersections between the  $-N_{is}$  circle and  $1/G(\sigma + j\omega_s)$ . One of these must be, due to the steady-state condition, at  $N_{is} = N$ , that is at  $\sigma = 0$ ,  $\gamma = 90^\circ$ . This solution corresponds to a phase shift of the steady-state oscillation.

#### Non-synchronous Case

The input to the non-linearity is now assumed to be

$$x = a e^{ip} + \Delta a e^{iq} \quad (23)$$

In this case the phase angle between the steady-state oscillation and the perturbation varies continuously. Hence,  $N_i$  for the non-synchronous case ( $N_{ins}$ ) is obtained<sup>4</sup> from  $N_{is}$  by averaging over the phase angle  $\gamma$ .

$$N_{ins} = N + \frac{a}{2} \cdot \frac{dN}{da} \quad (24)$$

#### $N_i$ for the Ideal Relay

Applying the above equations it is found that  $N_i$  for the ideal relay is given by

$$N_{is} = \frac{N}{2} (1 - e^{-2j\gamma}) \quad (25)$$

$$N_{ins} = \frac{N}{2}$$

### References

- 1 WEST, J. C., DOUCE, J. H., and LIVESLEY, R. K. The dual input describing function and its use in the analysis of nonlinear feedback systems. *Proc. Inst. elect. Engrs B* 103 (1956) 463
- 2 WEST, J. C. *Analytical Techniques for Non-Linear Control Systems* 1960. London; English Universities Press
- 3 BONENN, Z. Stability of forced oscillations in non-linear feedback Systems. *Inst. Radio Engrs Trans. on Automatic Control*, PGAC-6, Dec. (1958) 109
- 4 BONENN, Z. Relay amplifiers, *Ph. D. Thesis*, University of Cambridge (1960)

- <sup>5</sup> CODDINGTON, E. A., and LEVINSON, N. *Theory of Ordinary Differential Equations*. Chaps. 13, 14, 1955. New York; McGraw-Hill
- <sup>6</sup> TSYPKIN, JA. Z. *Theorie der Relaisysteme der Automatischen Regelung*. (Russian transl.) 1958. München; Oldenbourg
- <sup>7</sup> PIAGGIO, H. T. H. *Differential Equations*. p. 32. 1954. Bell
- <sup>8</sup> BONENN, Z. *The Incremental Input Describing Function and its Applications*, in the press

- <sup>9</sup> MACCOLL, L. A. *Fundamental Theory of Servomechanisms*. pp. 78-87. 1945. New York; Van Nostrand
- <sup>10</sup> GRENSTED, P. E. W. *The Frequency Response Analysis of Nonlinear Systems*. *Inst. elect. Engrs Monogr. No. 126* (1955)
- <sup>11</sup> GRENSTED, P. E. W. The transient response of non-linear control systems. *Ph. D. Thesis*, University of Cambridge, (1956)
- <sup>12</sup> COSGRIFF *Non-Linear Control Systems*. Chap. 9-5. 1958. New York; McGraw-Hill

## DISCUSSION

Z. BONENN, *Scientific Department, Ministry of Defence, P.O.B. 7063 Hakirya, Tel-Aviv, Israel*

## Author's Opening Remarks

## Nature of the Approximation

Further work was carried out to verify more closely the approximate solution presented in the paper. For an odd non-linearity the steady-state solution will have components at  $\omega_s$ ,  $3\omega_s$ ,  $5\omega_s$ ... Using the usual describing function approximation components at  $\omega \geq 3\omega_s$  are neglected in the steady solution. When the incremental input is added to the steady-state oscillation

$$x = a \sin \omega_s t + \Delta a \sin (\omega t + \gamma)$$

Output components will appear at  $\omega$  and  $2n\omega_s \pm \omega$  (odd non-linearity)  $n=1, 2, \dots$

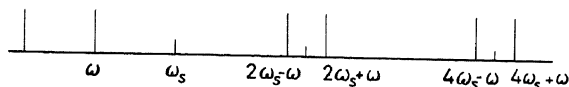


Figure A

Hence, the general form of a solution of eqn (2) is

$$\Delta x = e^{\sigma t} \left[ \Delta a_0 \sin (\omega t + \gamma_0) + \sum_{n=1}^{\infty} \{ \Delta a_n \sin [(2n\omega_s + \omega)t + \gamma_n] + \Delta a_n \sin [(2n\omega_s - \omega)t + \gamma_n] \} \right]$$

In line with the usual approximation we may neglect components at  $4\omega_s \pm \omega$  and above. Only the components at  $2\omega_s \pm \omega$  must be investigated. In the synchronous case  $\omega = \omega_s$  and  $2\omega_s \pm \omega \rightarrow \omega_s$ ,  $3\omega_s$ . The component at  $\omega_s$  resulting from  $2\omega_s - \omega_s$  is taken into account in  $N_{is}$ —the synchronous incremental describing function, indeed it is this component which makes  $N_{is}$  complex and of circular form. The component at  $3\omega_s$  is neglected as before.

The other solution in our example is mainly aperiodic and of the form

$$\Delta x = e^{\sigma t} \Delta a_0 [1 + \varepsilon_1 \sin (2\omega_s t + \gamma_1) + \dots] \omega = 0$$

Neglecting  $2\omega_s$  ( $\varepsilon_1 \approx 0$ ) and using eqn (13)

$$1 + N_{is} G(\sigma) = 0$$

we found  $\sigma = -10.55$

The second approximation is obtained by solving

$$1 + \left[ N_{is} + \frac{\varepsilon_1}{2} a \left( -\frac{d \operatorname{Re} N}{da} \sin \gamma_1 + \frac{d \operatorname{Im} N}{da} \cos \gamma_1 \right) \right] G(\sigma) = 0$$

$$1 + \left[ N_{is} - \frac{1}{\varepsilon_1} j a \frac{dN}{da} e^{-j\gamma_1} \right] \cdot G(\sigma + j 2\omega_s) = 0$$

After one iteration we find

$$\Delta x = \Delta a_0 \cdot e^{-10.5t} [1 - 0.12 \sin (2\omega_s t + 61^\circ)]$$

Thus we see that the approximation is well justified whenever the usual describing function approach is valid.

The validity of this approach may also be seen by comparison with the work of Dr. Gelb on limit cycling control systems published recently in *Trans. I.E.E.E.* April 1963. Dr. Gelb derives heuristically a second-order differential equation for the dynamics of the limit cycle, from which he calculates transient behaviour of the steady-state oscillation under test. Checking experimentally the system shown in *Figure B*, he found good agreement for small  $\zeta$ . However, for large  $\zeta$  there is a considerable difference between theory and experiment. For example, for

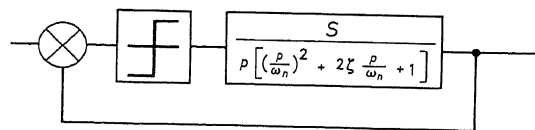


Figure B

Dr. Gelb's results are; Theoretical  $\frac{\tau}{T_0} = 3.22$

Experimental  $\frac{\tau}{T_0} \approx 2.25$

Our method gives;  $\frac{\tau}{T_0} = 2.36$

R. J. KOCHENBURGER, *University of Connecticut, Storrs, Connecticut, U.S.A.*

Dr. Bonenn has contributed much to the concept of the incremental describing function through this and his earlier papers. As he illustrates in the example, the method is an ingenious and useful one when investigating the degree of stability associated with small excursions from steady-state conditions and when dealing with an oscillating or limit-cycling type of system.

I attempted to extend, on my own, the author's technique to either what probably is the more usual situation when a non-oscillating steady-state response is the goal, or to the case of large excursions. I can report no success when using this particular technique. Could the author make some suggestions? For the example he treats, how would he consider a contactor with an inactive zone large enough to prevent oscillation? The best technique I was able to find for the problem just mentioned, was resorting to the usual describing function plot and superposing this on a family of loci of the function

$$\tilde{G}^1(-\varphi\omega_0 + j\sqrt{1-\varphi^2}\omega_0)$$

A plot then appears as shown in *Figure A*. When operating points for negative values of  $\varphi$  exist for the lower amplitudes, as they do in the example, a limit cycle is indicated. The time required to settle to such a steady-state condition, or the settling time for non-oscillating systems, may be reasonably approximated.



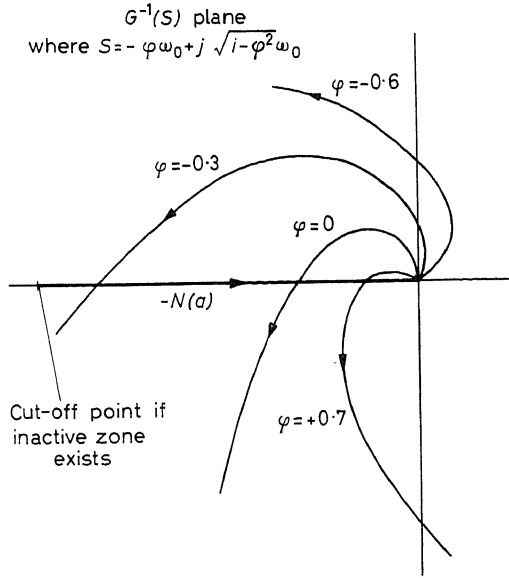


Figure A

In any event, comparisons drawn between various modifications of the describing function technique are meaningless unless accompanied by actual response examples on real systems or their analogue simulation. As a general suggestion for sessions on this subject, I would like to propose such a type of verification whenever it is possible. It is likely that such a procedure would have demonstrated the validity of Dr. Bonenn's interesting technique when applied to many practical problems.

Z. BONENN, *in reply*

I thank Professor Kochenburger for his interesting discussion and quite agree that the method is valid only for *small* excursions from the steady-state oscillation. I am afraid I can offer no solution for large excursions or transient conditions.

Dr. Grensted (Reference 11 of my paper) attacked this problem, but his approach becomes very complicated and difficult to apply for systems above second order.

I hope that my previous remarks on the accuracy of the approximation and the comparison with Dr. Gelb's experimental work demonstrates the validity of this technique.

J. L. DOUCE, *The Queen's University, Belfast, Northern Ireland*

The method is limited by the restriction that not only the perturbation but also the response to the perturbation must be small. This means that the complete response cannot be determined if the final state differs appreciably from the initial state in amplitude or frequency. Practical applications require an investigation of the stability of forced oscillations considered in the paper. This investigation must consider the possibility of the jump effect and the occurrence of subharmonic oscillations. Equation (10) can be used for this purpose. Does the author consider that eqn (9) has any advantage in this application?

I would suggest that a more usual definition of the describing function is the negative reciprocal of the author's  $N$ . The figure corresponding to Figure 4 now consists of the open-loop frequency response  $G(j\omega)$  and  $-1/N$ . I found the labelling of this figure confusing, due to this consideration.

The general expression for the incremental describing function, eqn (22), was, so far as I know, first published in 1955.

Z. BONENN, *in reply*

I quite agree with Dr. Douce that for calculating the jump region in the case of forced oscillations eqn (10), as shown by Professor West, should be used.

As for using  $N$  or  $-1/N$ , this is a question of taste and convenience. I use both.

I shall be grateful to Dr. Douce if he will give the 1955 reference to eqn (22). This previously was unknown to me.

T. F. JOOS DE TER BEERST, *Université Lovanium, B.P. 214, Leopoldville XI, Congo, Africa*

I should like to have some more details about the use of this theory when the non-linear function is piece-wise continuous and, in particular, of the use of the variational equation to establish the stability even in the case when  $f(x)$  is not everywhere continuous.

I would also like to point out that in some cases, the perturbation should have two terms instead of one, as stated in eqn (4). This can be shown using Floquet's theory, as will be shown in a paper to be published soon. These cases arise when the non-linear feedback system is able to oscillate freely without external excitation. If one studies the stability of this free oscillation, or of an externally excited oscillation at a frequency near that of the free oscillation, then the perturbation should be written

$$\Delta x = \Delta a e^{(\sigma + j\omega)t} + \Delta b e^{(\sigma - j\omega)t} \quad (1)$$

where  $\Delta a$  and  $\Delta b$  are two complex quantities. The phase-shift  $\gamma$  of eqn (4) of the paper has been included directly in  $\Delta a$ .

That there should be a second term in  $\Delta x$  can be shown as follows. Among all terms at the output of the non-linearity, there will be one at a frequency  $-\omega$ , and this term is not filtered out, if  $|\sigma|$  is small, which is implicitly admitted by the author since the term with frequency  $\omega$  is not filtered out. Using  $\Delta x$  in (1) as perturbation, eqn (6) becomes

$$\Delta a e^{j\omega t} + \Delta b e^{-j\omega t} + G(D + \sigma) \left( \frac{df}{dx} \right)_{x=x_{ss}} [\Delta a e^{j\omega t} + \Delta b e^{-j\omega t}] = 0 \quad (2)$$

One develops  $\left( \frac{df}{dx} \right)_{x=x_{ss}}$  in a Fourier series and obtains terms at frequencies  $0, 2\omega$  and  $-2\omega$ , and other terms

$$\left( \frac{df}{dx} \right)_{x=x_{ss}} = a_0 + a_2 e^{2j\omega t} + a_2 e^{-2j\omega t} + \dots$$

Replacing in (2)

$$\Delta a e^{j\omega t} + \Delta b e^{-j\omega t} + G(D + \sigma) [(a_0 \cdot \Delta a + a_2 \cdot \Delta b) e^{j\omega t} + (a_0 \cdot \Delta b + a_2 \cdot \Delta a) e^{-j\omega t} + \dots] = 0$$

where the neglected terms have a frequency different from  $\pm \omega$ . Grouping terms, equating separately the coefficients of  $e^{j\omega t}$  and  $e^{-j\omega t}$  to zero

$$\begin{cases} \Delta a + G(\sigma + j\omega) [a_0 \cdot \Delta a + a_2 \cdot \Delta b] = 0 \\ \Delta b + G(\sigma - j\omega) [a_0 \cdot \Delta b + a_2 \cdot \Delta a] = 0 \end{cases}$$

For this system to have a solution in  $\Delta a, \Delta b$  different from zero, the following determinant should be zero.

$$\begin{vmatrix} 1 + a_0 G(\sigma + j\omega) & a_2 G(\sigma + j\omega) \\ a_2 G(\sigma - j\omega) & 1 + a_0 G(\sigma - j\omega) \end{vmatrix} = 0 \quad (3)$$

This equation replaces eqn (9) of the paper. In addition it can be shown by Floquet's theory that  $\sigma$  should have a small magnitude and that it

is therefore allowed to develop  $G(\sigma + j\omega)$  and  $G(\sigma - j\omega)$  in a Taylor series keeping only the first two terms

$$G(\sigma + j\omega) = G(j\omega) + \sigma \left( \frac{dG(u)}{du} \right)_{u=j\omega} + \dots$$

$$G(\sigma - j\omega) = G(-j\omega) + \sigma \left( \frac{dG(u)}{du} \right)_{u=-j\omega} + \dots$$

Thus eqn (3) is always of the second degree in  $\sigma$ , and at most two stability conditions have to be met to ensure that the real part of  $\sigma$  should be negative.

#### References

- <sup>1</sup> BELEVITCH, V. Perturbations dans les systèmes non-linéaires filtrés. Applications à la théorie des oscillateurs. *Revue HF*, 2, No. 12 (1954)
- <sup>2</sup> BELEVITCH, V. *Théorie des circuits non-linéaires en régime alternatif*. 1959. Louvain; Uystpruyst
- <sup>3</sup> ROUCHE, N., NEIRYNCK, J. and JOOS DE TER BEERST, T. Théorie des circuits non-linéaires filtrés. *Université Lovanium, Rapport de recherches du laboratoire d'électricité*, No. 1. 1960. Leopoldville

Z. BONENN, *in reply*

I thank Mr. Joos de ter Beerst for his interesting discussion. However, I cannot agree with his remarks. Indeed, I cannot see why the two terms in his equation (1) could not be combined into one term at frequency  $\omega$ .

It seems to me that this difference of opinion arises as follows. Going back to the general form of the perturbation we observe (see

opening remarks) that as  $\omega \rightarrow \omega_s$  the perturbation consists of two terms at  $\omega$  and  $2\omega_s - \omega$  or if we wish  $\omega_s - \Delta\omega$  and  $\omega_s + \Delta\omega$ . When  $\Delta\omega \rightarrow 0$  these two terms coalesce, it is this coming together of the two terms which creates the synchronous incremental describing function which is complex and of circular form. Hence, when using our eqn (9) with  $N_{is}$  we take care of both terms.

The situation is quite different when  $\Delta\omega \neq 0$ . This is the difficult 'almost synchronous' case. Mr. Joos de ter Beerst's equations [before (3)] represent this case. Noting that

$$a_0 = N_{ins}; \quad a_2 = -\frac{a}{2} \cdot \frac{dN}{da}; \quad a_{-2} = -\frac{a}{2} \cdot \frac{dN^*}{da}$$

and that his  $\Delta a$ ,  $\Delta b$  are complex, his equations correspond to our equations for the 'almost synchronous' case.

$$1 + \left[ N_{ins} + \varepsilon_1 \cdot \frac{a}{2} \cdot \frac{dN^*}{da} \cdot e^{-j(\gamma_1 + \gamma_2)} \right] G[\sigma + j(\omega_s - \Delta\omega)] = 0$$

$$1 + \left[ N_{ins} + \frac{1}{\varepsilon_1} \cdot \frac{a}{2} \cdot \frac{dN}{da} \cdot e^{-j(\gamma_1 + \gamma_2)} \right] G[\sigma + j(\omega_s + \Delta\omega)] = 0$$

When one inspects these equations it turns out that it is extremely difficult to fulfil all the necessary conditions for their simultaneous solution. Indeed, I have not yet succeeded in finding any system which has such a solution. Moreover, it is easy to see that such solutions are impossible in second- and third-order systems because there must be at least two synchronous solutions. I have tried to obtain an almost synchronous solution in a fourth-order system, but without success. Here I obtained either two synchronous and two non-synchronous solutions or four synchronous solutions.

# Predictive Control of an On-Off System with Two Control Variables

A. J. ADEY, J. F. COALES and J. A. STILES

## Summary

This paper studies the optimum control of multi-variable systems by considering the behaviour of the two-variable on-off system described by the equation

$$\ddot{x} + A\dot{x} = u$$

where  $u_1, u_2 = \pm 1$ .

By the application of Pontryagin's maximum principle, it is shown how the number of switchings and the drive ratio  $u_1/u_2$ , for an optimum trajectory (minimum settling time criterion), depend on the nature of the matrix  $A$ . For the case when the interaction terms  $a_{12}$  and  $a_{21}$  have the same sign, the maximum number of switchings for an optimum trajectory is three, and the drive ratio  $u_1/u_2$  near the origin of the error phase space has the same sign as  $a_{12}$  and  $a_{21}$ .

A predictive controller described in this paper finds the three switching times by an iterative process which involves the application of logical rules to the predicted behaviour of the system, computed in a fast analogue model. This realizable controller is optimum when the two-by-two matrix  $A$  has  $a_{12} a_{21} \geq 0$ ; and will also control systems where  $a_{12} a_{21} < 0$ , although not optimally. It is also shown that the problem of hunting may be overcome by confining the hunting to the fast model alone.

## Sommaire

La présente communication étudie la commande optimale de systèmes à variables multiples en considérant le comportement du système par Tout ou Rien à deux variables, décrit par l'équation  $\ddot{x} + A\dot{x} = u$  (où  $u_1, u_2 = \pm 1$ ).

L'application du principe du maximum de Pontryagin montre la manière dont le nombre de commutations et le rapport  $u_1/u_2$  pour la trajectoire optimale (critère du temps d'établissement minimum), dépendent de la nature de la matrice  $A$ . Lorsque les termes d'interaction  $a_{12}$  et  $a_{21}$  possèdent le même signe, le nombre maximum de commutations sur une trajectoire optimale s'élève à trois et le rapport  $u_1/u_2$ , au voisinage de l'origine de l'espace de phases de l'erreur, possède le même signe que  $a_{12}$  et  $a_{21}$ .

Un régulateur à prédiction, décrit dans la présente communication, trouve les trois moments de commutation au moyen d'un processus itératif qui applique des règles logiques au comportement prédit du système, calculé sur un modèle analogue rapide. Ce régulateur réalisable est optimal lorsque la matrice deux-par-deux  $A$  vérifie la condition  $a_{12} a_{21} \geq 0$ ; il peut également commander — quoique non-optimalement — des systèmes pour lesquels  $a_{12} a_{21} < 0$ . Il est également montré que le problème d'auto-oscillations peut être résolu en limitant ces auto-oscillations au seul modèle rapide.

## Zusammenfassung

Der Aufsatz untersucht die optimale Regelung von Mehrfachsystemen, indem er das Verhalten eines Zweipunktreglers für zwei Variable, das durch die Gleichung  $\ddot{x} + \dot{x}A = u$ , wobei  $u_1, u_2 = \pm 1$ , beschrieben wird, betrachtet.

Die Anwendung des Pontryaginischen Maximumprinzips zeigt, wie die Anzahl der Umschaltungen und das Antriebsverhältnis  $u_1/u_2$  für eine optimale Trajektorie (Kriterium der minimalen Einstellzeit) von der Beschaffenheit der Matrix  $A$  abhängt. Für den Fall, daß die Kopplungsglieder  $a_{12}$  und  $a_{21}$  das gleiche Vorzeichen haben, ist die maximale Anzahl der Umschaltungen für eine optimale Trajektorie drei und das Antriebsverhältnis  $u_1/u_2$  hat in der Nähe des Ursprungs des (Fehler-)Phasenraumes das gleiche Vorzeichen wie  $a_{12}$  und  $a_{21}$ .

Ein in dieser Arbeit beschriebener Regler mit Vorhersageeigenschaften ermittelt die drei Schaltzeitpunkte durch einen iterativen Vorgang, der die logischen Gesetze auf das Vorhersageverhalten des Systems anwendet, das ein schnelles Analogmodell berechnet. Der realisierbare Regler arbeitet dann optimal, wenn in der quadratischen Matrix (2. Ordnung)  $A$   $a_{12} a_{21} \geq 0$  ist; es tritt auch eine Regelung von Systemen mit  $a_{12} a_{21} < 0$  auf, diese ist jedoch nicht optimal. Es wird gezeigt, daß sich das Problem der stabilen Dauerschwingungen (hunting) ebenfalls dadurch bewältigen läßt, daß man die stabilen Dauerschwingungen auf das schnelle Modell beschränkt.

## Introduction

The practical design of control systems with more than one variable has generally used traditional linear methods. These methods often involve nullifying the effect of interaction between the variables, which normally increases the complexity of the controller. However, in all practical systems either the input force (or torque) of the plant motor or the input power (or energy) to the plant is limited, and in many such cases it is known that the best control possible, in some sense, is obtained by using an on-off controller. Even if on-off control is not optimum, it will nevertheless usually be nearly so.

When such controllers were first considered for single-variable systems, they were designed to operate by generating a switching function. However, in all but the simplest cases, the complexity involved in this is prohibitive. For this reason the method of predictive control was developed in the early 1950s<sup>1</sup>.

In its first form this controller simulated repetitively the future behaviour of the system on a fast time scale. The information provided by this simulation enabled the switching time giving the desired system behaviour to be found by a process of iteration. Subsequently, by using the general properties of the switching function, it has become possible to determine the correct drive direction by a single simulation<sup>2</sup>. A further development of these ideas is necessary for a satisfactory solution of the two variable on-off control problem.

The two-variable plant considered here is that described by the equation

$$\ddot{x} + A\dot{x} = u \quad (1)$$

where

$$x = \begin{pmatrix} x_1 \\ x_2 \end{pmatrix}, \quad u = \begin{pmatrix} u_1 \\ u_2 \end{pmatrix}, \quad A = \begin{pmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{pmatrix}$$

and  $u$  is constrained (Figure 1).

The more general system described by the equation

$$\ddot{y} + C\dot{y} = Bu$$

with  $B$  non-singular may be reduced to this form by putting

$$y = Bx, \quad A = B^{-1}CB$$

Such a system, i.e. second order with only inertia and viscous friction terms, forms the basis of many practical systems. Furthermore, the complexity of the controller increases rapidly with the number of variables and so it is necessary to understand the two-variable case before proceeding to more variables.

This paper shows that, using minimum settling time as the performance criterion, the maximum number of switches for an optimum trajectory can be three, four or infinite, depending on the nature of the matrix  $A$ . If the interaction terms  $a_{12}$  and  $a_{21}$  have the same sign, this number is three and the problem becomes simpler to solve. A practical method is given for controlling a system of this type.

Moreover, it is likely that, as in the single-variable case<sup>1</sup>, the same technique (of using three switches) can be used to control satisfactorily, though not optimally, this plant for any  $A$ , or to control two-variable systems with non-linearities or of higher order; even when the desired outputs are slowly varying random signals. Furthermore, following Chestnut<sup>2</sup>, it may only be necessary to use an approximate simulation of the plant, e.g. a second-order one. It is also shown how the problem of hunting may be solved and so the controller described in this paper is likely to have wide application.

### The Number of Switches

For the system (1) with  $|u_i| \leq 1$ , and desired behaviour that the position and velocity should become zero in minimum time, it is well known how to 'trace backwards' optimum trajectories from the origin of the phase space<sup>3</sup>.

For 'back tracing' the system equation becomes

$$\ddot{x} - A\dot{x} = u$$

The adjoint equation is

$$\ddot{p} + A'\dot{p} = 0$$

This is integrated to give

$$\dot{p} + A'p = \beta \quad (2)$$

where  $\beta$  is a constant vector.

The optimum drive is given by

$$u = \text{sgn } p \quad (3)$$

Hence by choosing different initial condition vectors  $p(0) = \alpha$  and  $\dot{p}(0) = \beta - A'\alpha$  for the adjoint equation, different optimum trajectories are produced.

To determine the complete optimum drive  $u(t)$  from eqns (2) and (3) one needs to know, first, when the zeros of the components of  $p$  occur, and, second, what the drives are for some given time, say  $t = 0$ . These two questions are considered in turn.

### Zeros of $p$

The explicit solution of eqn (2) depends on the nature of the eigenvalues  $\lambda_1, \lambda_2$  of  $A$ ; and on this solution depends the possible number of zeros of the components of  $p$ , i.e. switches. It is helpful to classify  $A$  by its cross-terms  $a_{12}$  and  $a_{21}$  as follows.

(i) If  $a_{12}a_{21} < -\{(a_{11} - a_{22})/2\}^2$ ,  $\lambda_1$  and  $\lambda_2$  are complex, and there is no upper bound to the possible number of switches.

(ii) If  $-\{(a_{11} - a_{22})/2\}^2 \leq a_{12}a_{21} < 0$ , the number of switches can never exceed four (two in each component of  $u$ ).

(iii) If  $0 \leq a_{12}a_{21}$ , the number of switches can never exceed three (two in one component of  $u$  and one in the other).

These results do not apply to a small region of the state space where the system is singular i.e. 'non-normal'. This region contains all points for which the optimum drive has no switch in one of the components of  $u$ .\*

In order to prove (ii) and (iii), which correspond to  $\lambda_1$  and  $\lambda_2$  real, only the case in which  $\lambda_1$  and  $\lambda_2$  are distinct and non-zero will be considered. The cases of equal eigenvalues or one zero eigenvalue may be similarly dealt with.

In this case, the solution of eqn (2) may be written

$$p = (A'^{-1})\beta + \frac{e^{-\lambda_1 t}}{\lambda_1 - \lambda_2}(A' - \lambda_2 I)\left(\alpha - \frac{\beta}{\lambda_1}\right) - \frac{e^{-\lambda_2 t}}{\lambda_1 - \lambda_2}(A' - \lambda_1 I)\left(\alpha - \frac{\beta}{\lambda_2}\right) \quad (4)$$

which shows that neither component of  $p$  can have more than two zeros.

If a back-traced optimum trajectory contains four switches, at  $t_1$  and  $t_2$  in  $u_1$ , and  $s_1$  and  $s_2$  in  $u_2$ , there must be values of  $\alpha$  and  $\beta$  satisfying

$$p_1(t_1) = p_1(t_2) = p_2(s_1) = p_2(s_2) = 0$$

These four equations, linear and homogeneous in the four components of  $\alpha$  and  $\beta$ , have a non-trivial solution if, and only if, the determinant of the coefficients vanishes. This condition reduces to

$$\frac{a_{11} - \lambda_1}{a_{11} - \lambda_2} = \frac{f(t_1, t_2)}{f(s_1, s_2)} \quad (5)$$

where

$$f(x, y) = \frac{e^{-\lambda_1 x} - e^{-\lambda_1 y}}{e^{-\lambda_2 x} - e^{-\lambda_2 y}}$$

This is therefore the necessary and sufficient condition that  $t_1, t_2, s_1, s_2$  should be switching times on some optimum trajectory.

If  $\lambda_1 > \lambda_2$ ,  $f(x, y)$  is non-vanishing, continuous, and tends to zero along the line  $x = y$  as  $x \rightarrow \infty$ . Thus eqn (5), being the quotient of two values of  $f(x, y)$ , can take any positive value; i.e. by suitably choosing  $t_1, t_2, s_1, s_2$ , eqn (5) will always be satisfied provided  $a_{11} - \lambda_1$  and  $a_{11} - \lambda_2$  have the same sign. If, however, they have opposite signs, or if one is zero, eqn (5) can never be satisfied. Substituting

$$\lambda_1, \lambda_2 = \frac{1}{2}[a_{11} + a_{22} \pm \{(a_{11} - a_{22})^2 + 4a_{12}a_{21}\}^{1/2}] \quad (6)$$

(ii) and (iii) are proved.

### Drive Directions

The drive ratio  $u_2/u_1$  depends on the solution of eqn (4) and the cross-coupling terms  $a_{12}$  and  $a_{21}$ .

If  $0 \leq a_{12}a_{21}$ , viz. a type (iii) system, this ratio always has the same sign at the origin of the state space, i.e. at  $t = 0$  for 'back tracing'.

\* This region, as well as the boundaries in state space for the various types of drive, has been extensively analysed since the paper was written.

In order to prove this, consider the case of two switches at times  $t_1$  and  $t_2$  in  $u_1$ , and one switch at time  $s_1$  in  $u_2$ . The proof is similar when there are two switches in  $u_2$  and one in  $u_1$ .

If  $a_{21} \neq 0$  the ratio  $u_2/u_1$  at the origin may be written

$$\frac{\alpha_2}{\alpha_1} = \frac{(a_{11} - \lambda_1) h_1 (g_{21} - g_{22}) - (a_{11} - \lambda_2) h_2 (g_{11} - g_{12})}{a_{21} (g_{11} g_{22} - g_{12} g_{21})}$$

where

$$g_{ij} = 1 - e^{-\lambda_i t_j} \quad \text{and} \quad h_i = 1 - e^{-\lambda_i s_i}$$

Since  $g_{ij}/g_{2j} = f(0, t_j)$  is monotonic in  $t_j$  the denominator can never vanish. Thus, as the ratio is a continuous function of  $t_1, t_2, s_1$ , if its sign changes it must vanish for some positive  $t_1, t_2, s_1$ . This means  $\alpha_2 = 0$ , i.e. that there is a fourth switch at the origin, which has just been shown to be impossible for a type (iii) system.

To see what this sign is, suppose  $\lambda_1 > \lambda_2$  and let  $t_1 \rightarrow \infty$ . The ratio then becomes

$$\frac{(a_{11} - \lambda_1) h_1 e^{-\lambda_2 t_2} - (a_{11} - \lambda_2) h_2 e^{-\lambda_1 t_2}}{a_{21} (e^{-\lambda_1 t_2} - e^{-\lambda_2 t_2})} \quad \text{if } \lambda_2 > 0$$

or

$$-\frac{(a_{11} - \lambda_1) h_1}{a_{21} g_{12}} \quad \text{if } \lambda_2 < 0$$

Since for  $a_{12} a_{21} \geq 0$ , by eqn (6),  $\lambda_1 > \lambda_2$  implies  $\lambda_1 > a_{11}$  (and in this case  $a_{11} - \lambda_1$  and  $a_{11} - \lambda_2$  have opposite signs), both these ratios have the same sign as  $a_{21}$ ; i.e. the drive ratio at the origin has the same sign as the cross-coupling terms.

For a forward-time optimum trajectory containing three switches, the starting drive ratio is therefore opposite in sign to the cross-coupling terms, provided  $a_{12} a_{21} \geq 0$ .

On the other hand, if  $a_{12} a_{21} < 0$  the drive ratio at the origin of the state space may be different for different trajectories.

The controller described below has been developed for the case  $a_{12} a_{21} \geq 0$ . These results considerably simplify its construction: for, in searching for the optimum trajectory three times may be taken as parameters; and, of the four possible starting drives, only two are permissible.

### The Control Strategy

A logical strategy has been developed for determining the correct drive vector to control the plant described by eqn (1). So far only the case with all the coefficients  $a_{11}, a_{12}, a_{21}$  and  $a_{22}$  positive has been considered. The general layout of the plant and controller is given in Figures 1 and 2. Figure 1 shows the system with  $S$  as the Laplace operator, while Figure 2 shows the controller in which  $G$  gates the initial conditions for the model and  $s$  is the Laplace operator with respect to fast-model time.

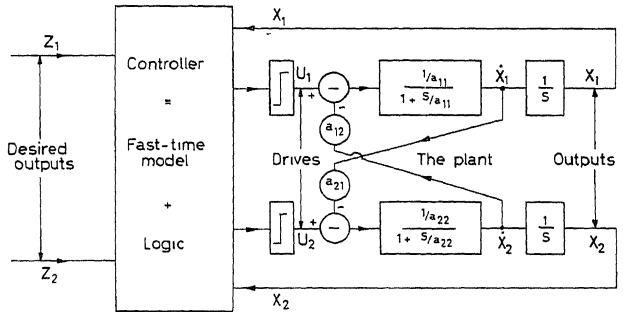


Figure 1

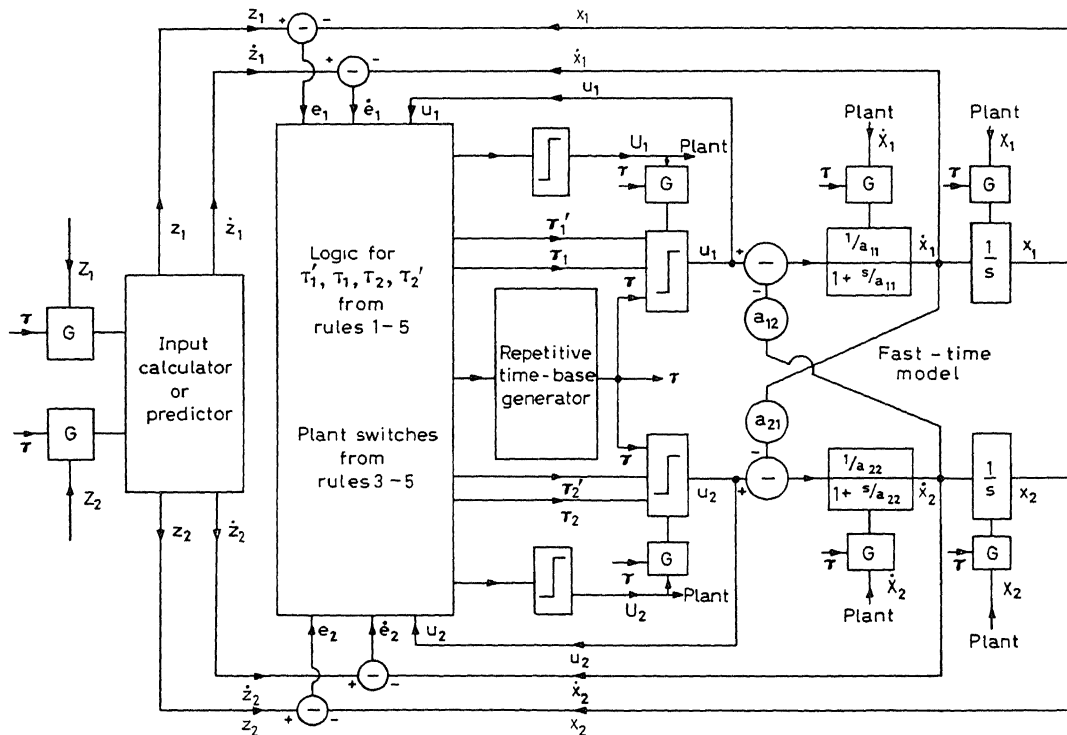


Figure 2

two projections  $P_1$  and  $P_2$  of the trajectory on the phase plane of error and its derivative for each of the two variables. This is satisfactory because, due to the nature of eqn (1), each output variable is affected most strongly by the corresponding drive; and in fact the projections  $P_1$  and  $P_2$  resemble the trajectories produced by single-variable second-order systems, although somewhat distorted by interaction.

These projections are generated repetitively on a fast time scale by an input calculator and an analogue model of the plant. The optimum switching times are found in this fast model from arbitrary values by a process of iteration which uses the shape and position of the projections in one time sweep to find better switching times for the next time sweep, and when necessary to alter the drive directions.

In general there are three switches, say for example two in drive  $u_1$  and one in drive  $u_2$ . The shape of the projection  $P_1$  indicates that one switch in drive  $u_1$ , called the main switch, may be adjusted to make the projection  $P_1$  pass through the origin of its phase plane. Similarly the main switch in drive  $u_2$  may be adjusted to make the projection  $P_2$  pass through the origin of its phase plane. It is also apparent that the other switch in drive  $u_1$ , called the supplementary switch, may be adjusted to make the two projections reach their respective origins at the same instant of time.

As there are two switches in drive  $u_1$ , these may be adjusted to make the trajectory pass through the origin of the phase space, no matter which direction the drive  $u_1$  has initially. The initial direction which gives the shorter settling time is therefore chosen and this is found by experiment to be that which gives the initial drive ratio  $u_1(0)/u_2(0) = -1$ . This agrees with the earlier result (here  $a_{12}$  and  $a_{21}$  are both positive).

The two main switching times can be adjusted simultaneously because their interaction merely slows their convergence to steady values. However, interaction can prevent the supplementary switching time being adjusted to its proper value simultaneously with the main switching times; so it is necessary to keep the supplementary switching time fixed while the main switching times are converging to steady values. Whenever these values have been reached, the supplementary switching time is adjusted; after which the main switching times converge to new values. This process is repeated until the supplementary switching time has also arrived at a steady value, when the model trajectory will be the optimum one, except in the small singular, i.e. 'non-normal', region of the state space. As no special provision is made for this region, the control here is good but not optimum.

As the directions of the plant drives are the same as the initial drive directions in the model, this system will cause the plant drives to be optimum except possibly during the iterative process in the model. This however only occurs after a significant change in the desired output and lasts for only a short time (up to 0.8 sec in this system where each time sweep lasts a maximum of 0.04 sec and the basic plant time constant is 10 sec).

This iterative process is governed by the rules set out below.

### The Detailed Logical Rules

In the model, time  $\tau$  is measured during each time sweep with  $\tau$  set to zero at the beginning of the trajectory. Each time sweep lasts until  $\tau = \tau_f$ , a time greater than any possible settling time;

and between each time sweep the initial conditions for the model and its drives, as well as those for the input calculator, are set equal to the present conditions in the plant. The logical rules for adjusting the switching times, etc. are applied during this resetting period.

At any instant, the controller operates in one of three possible modes. These are:

*Mode 1.* Here the supplementary switch occurs in the drive  $u_1$  at time  $\tau'_1$ .

*Mode 2.* Here there is no supplementary switch.

*Mode 3.* Here the supplementary switch occurs in the drive  $u_2$  at time  $\tau'_2$ .

In all modes the main switches occur at  $\tau_1$  in the drive  $u_1$  and at  $\tau_2$  in the drive  $u_2$ . If a supplementary switch does not exist, the corresponding  $\tau'$  is zero in the calculations which follow, e.g. in Mode 1,  $\tau'_2 = 0$  and in Mode 2,  $\tau'_1 = \tau'_2 = 0$ .

Hereafter only Modes 1 and 2 are considered, because the rules for Mode 3 are found from those for Mode 1 by interchanging the suffices 1 and 2 throughout.

For convenience in the electronics, in Mode 1 the switches in the drive  $u_1$  are ordered so that  $\tau'_1 < \tau_1$ . Let:

$$e_1 = z_1 - x_1$$

where  $z$  is the desired output.

$$d_1(\tau) = -u_1(\tau) \cdot e_1(\tau)$$

$$m_1(\tau) = |e_1(\tau)| + k_1 |\dot{e}_1(\tau)|$$

$$q_1(\tau) = \text{sgn} \{-u_1(\tau) \cdot \dot{e}_1(\tau)\}$$

where  $k_1$  is a positive constant chosen by experiment.

In each time sweep, the following test is made during the interval  $\tau'_1 < \tau < \tau_1$  (Figure 3).

Is  $d_1 \geq \epsilon$  at the instant  $\tau_{a1}$  when  $q_1 = +1$  for the first time? ( $\epsilon$  is a small positive constant chosen by experiment.)

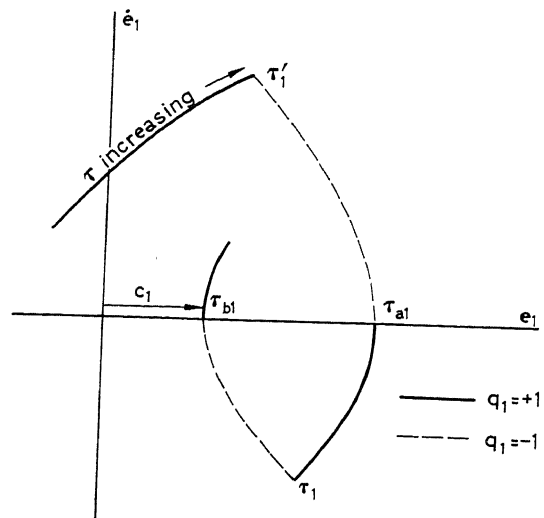


Figure 3

Depending on the result of this test, the following action is taken:

If yes:

- (i) Store the function  $c_1 = d_1(\tau_{a1})$  for use under Rule 1.
- (ii) Set the resetting time  $\tau_{r1} = \tau_{a1}$  (see note on p. 46).
- (iii) Apply Rule 3 after this time sweep.

If no:

Make the following test at the switching instant when  $\tau = \tau_1$ .

Does  $q_1$  change from  $-1$  to  $+1$ ?

If yes:

- (i) Store  $c_1 = m_1(\tau_1)$ .
- (ii) Set  $\tau_{r_1} = \tau_1$ .

If no:

Make the following test during the interval  $\tau_1 < \tau < \tau_f$ .

Does the instant  $\tau_{b_1}$  at which  $q_1 = +1$  for the first time exist?

If yes:

- (i) Store  $c_1 = d_1(\tau_{b_1})$ .
- (ii) Set  $\tau_{r_1} = \tau_{b_1}$ .
- (iii) Store the settling time  $\tau_{s_1} = \tau_{b_1}$  (see below).

If no:

- (i) Store  $c_1 = 0$ .
- (ii) Set  $\tau_{r1} = \tau_f$ .
- (iii) After the time sweep, find the new value for  $\tau_1$  by setting  $\tau_1$  to one tenth of its previous value.

Similar tests are performed on the second variable by changing the suffix 1 to 2 throughout the above.

If as a result of these tests  $\tau_{s1} = \tau_{b1}$ ,  $\tau_{s2} = \tau_{b2}$  and  $c_1 + c_2 < \gamma$  the difference between the settling times is sufficiently accurate for use in Rules 1 or 4, and the function  $c_{s1} = \tau_{s2} - \tau_{s1}$ . Otherwise  $c_{s1} = 0$ . ( $\gamma$  is a constant chosen by considering the stability of the  $\tau'_1$  and  $\tau_1$  control loops.)

After each time sweep the new values for the switching times are found by applying Rule 1. The other rules only apply occasionally, e.g. after a change in the desired output—Rule 2 checks that the drive ratio obeys the condition found earlier; Rule 3 changes the mode or the drive direction when the initial guess for this has been wrong; Rule 4 governs the change from Mode 2 to Mode 1 or 3 whenever this is necessary; Rule 5 gives the switches in the plant drive whenever  $\tau'_1$ ,  $\tau_1$  or  $\tau_2$  becomes zero.

*Rule 1.* Form new values for  $\tau_1$  and  $\tau_2$  by adding  $\mu_1 c_1$  to  $\tau_1$

and  $\mu_9 c_2$  to  $\tau_2$ . Also in Mode 1 form the new value for  $\tau'_1$  by adding  $\psi_1 c_{s1}$  to  $\tau_1$ .  $\mu_1, \mu_2, \psi_1$  are constants sufficiently small for the pulsed control loops for  $\tau_1, \tau_2, \tau'_1$  to be stable under all conditions (*Figure 4*).

The speed of convergence of  $\tau_1$  and  $\tau'_1$  may be considerably increased by making the following additional corrections: in Mode 1, whenever a change is made in  $\tau'_1$ , also add  $\eta_{s1}c_{s1} + \eta_{11}c_1 + \eta_{21}c_2$  to  $\tau_1$  and  $\eta_{11}c_1 + \eta_{21}c_2$  to  $\tau'_1$  where  $\eta_{s1}$  etc. are constants (positive or negative) chosen by experiment to maximize the speed of convergence of  $\tau_1$  and  $\tau'_1$ .

*Rule 2.* In Mode 1, if in the plant  $U_2/U_1 = \text{sgn } a_{21}$  [and in the previous time sweep in the model  $|d_2(\tau_{a_2})| \geq \varepsilon$ , change to Mode 2 and set  $\tau_1 = \mu_1 m_1(0)$ ].

*Rule 3.* If in the previous time sweep  $d_1(\tau_{\alpha_i}) \geq \varepsilon$  (see above), set  $\tau_1 = \tau'_1$  before applying Rule 1 and also, if in Mode 1 change to Mode 2, or if already in Mode 2 switch the plant drive  $U_1$ .

**Rule 4.** In Mode 2, if  $c_{s1} \geq \varepsilon'$  changeto Mode 1 before applying Rule 1 ( $\varepsilon'$  is a small positive constant chosen by experiment). Also if the plant drive ratio  $U_2/U_1 = \text{sgn } a_{21}$  correct this by switching the plant drive  $U_1$ . On the other hand, if already  $U_2/U_1 = -\text{sgn } a_{21}$  set  $\tau'_1 = \tau_1$  before applying Rule 1 and then set  $\tau_1 = \mu_1 m_1(0)$ .

Similarly if  $c_{s_2} (= -c_{s_1}) \geq \varepsilon'$  change to Mode 3 and apply the above logic with the suffices 1 and 2 interchanged.

*Rule 5.* After applying Rule 1,

- (i) if in Mode 1,  $\tau'_1 \leq 0$  switch the plant drive  $U_1$  and change to Mode 2;
- (ii) if in Mode 1,  $\tau_1 \leq 0$  change to Mode 2 and set  $\tau_1 = \mu_1 m_1(0)$ ;
- (iii) if in Mode 2,  $\tau_1 \leq 0$  switch the plant drive  $U_1$  and set  $\tau_1 = \mu_1 m_1(0)$ ;
- (iv) if in Mode 1 or 2,  $\tau_2 \leq 0$  switch the plant drive  $U_2$  and set  $\tau_2 = \mu_2 m_2(0)$ .

With an accurate model and input calculator, the accuracy of the control depends on the speed with which the searching process is carried out. This means that the model time scale should be about one thousandth of real time and that the length and number of time sweeps required should be minimized. Much of the complexity in the rules is due to this aim of using every piece of information available when computing new switching times, etc. Also a particular time sweep is ended as soon as all possible information has been obtained, as suggested by CHESTNUT<sup>2</sup>, by making the time sweep end at  $\tau_r$  where:

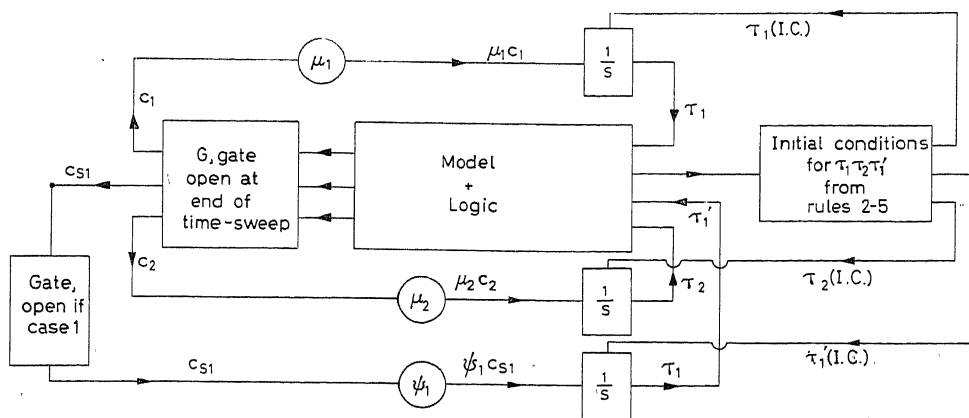


Figure 4

Note.  $\tau_r$  is the larger of  $\tau_{r1}$  and  $\tau_{r2}$ . To ensure the stability of the iterative process, the model drive  $u_1(\tau) = 0$  for  $\tau > \tau_{r1}$  and  $u_2(\tau) = 0$  for  $\tau > \tau_{r2}$ .

In practice, the optimum trajectory found in advance by the model and the actual trajectory followed by the plant will differ slightly due to noise and other inaccuracies. The resulting recomputing of the optimum trajectory by the model could often cause the plant drive to be switched several times with consequent increase in power consumption and relay wear. For this reason the model only computes a new trajectory when it has moved significantly away from the optimum one; the small quantities  $\varepsilon$  and  $\varepsilon'$  in the rules are chosen accordingly. These rules give an on-off system which inevitably hunts when the plant output reaches the desired output, resulting in unnecessary power dissipation. In general, providing the frequency of hunting is several octaves higher than the highest frequency which the plant is required to follow, the hunting may be smoothed out without loss of performance; and provided this is so the hunting in this system may be confined completely to the model by making the following changes in the controller:

(1) Replace the power relays for the plant by low-power relays followed by low-pass filters and power amplifiers. Similar low-pass filters must of course also be included in the fast model.

(2) If  $m_1(0) + m_2(0) < 3\varepsilon$  always go to Mode 2 and modify Rule 5 (iii) and (iv) to read as follows:

If in Mode 2,  $\tau_i \leq \varepsilon'$  switch the plant drive  $U_i$  and set  $\tau_i = \theta$  where  $i = 1, 2$ .  $\theta$  is a constant chosen so that the model makes about five time sweeps between each switch of one plant relay. This last system has the advantage over many on-off systems that it will operate satisfactorily with ramp desired outputs. Under these conditions the above rules cause the square wave output of the plant relay to have a mark to space ratio appropriate for following the desired ramp output.

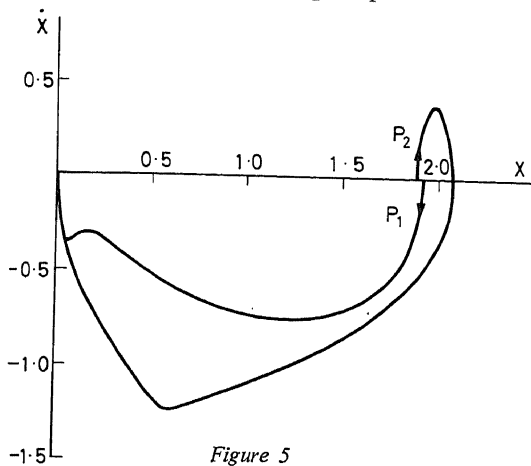


Figure 5

Figure 5 shows the projections on the output phase plane of a typical trajectory when the desired output  $Z$  is zero. Figure 6 shows the output, output rate and drive for the same trajectory, plotted against time.

### Conclusions

This paper presents a practical method for realizing the optimum controller for the two-variable system described by the differential equation  $\ddot{x} + A\dot{x} = u$ , where the terms  $a_{12}$  and  $a_{21}$  of the matrix  $A$  have the same sign. The controller is an

on-off one using predicted changeover, and it is shown how the problem of hunting can be solved. Although the system described here has been developed for determinate but varying desired outputs, it is made applicable to random desired outputs by substituting an input predictor for the input calculator.

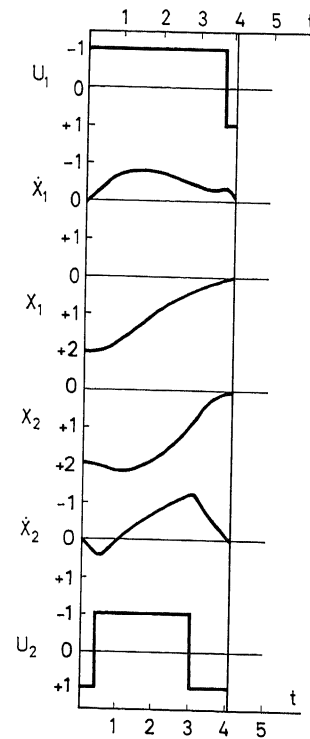


Figure 6

As this system solves the Pontryagin equations in under fifty iterations it has considerable advantages over other iterative methods of solution, e.g. hill climbing, which may require several hundred iterations. It is proposed to extend this method to cover the more difficult case of  $a_{12}$  and  $a_{21}$  having opposite signs. The optimum controller in this case will require more than three switches; so the advantages of such a complex optimum controller will be weighed against the simplicity of a non-optimum controller based on the present method, by comparing back-traced optimum trajectories with non-optimum three-switch trajectories.

Owing to the flexibility and logical nature of the rules governing this method, it is hoped that this type of controller will be applicable to a wide variety of two-variable plants and also that the logic may be extended to cover three, and eventually more, variable systems.

This research was supported by grants from the Department of Scientific and Industrial Research and the Commonwealth Scientific and Industrial Research Organisation (Australia).

### References

- COALES, J. F. and NOTON, A. R. M. An on-off servomechanism with predicted change-over. *Proc. Instn. Elect. Engrs* 103 (1955) 449
- CHESTNUT, H., SOLLECITO, W. E. and TROUTMAN, P. H. Predictive-control system application. *Trans. Amer. Inst. elect. Engrs* (1961) 128
- ROZENBERG, L. I. L. S. Pontryagin's maximum principle in the theory of optimum systems, part II. *Automat. Telemekh., Moscow* 20 (November 1959)



# A Method of Prediction for Non-stationary Processes and its Application to the Problem of Load Estimation

E. D. FARMER

## Summary

A method of predicting a non-stationary process is described. It is shown that, if the sample functions of the process are expanded in terms of the eigenfunctions of the Loeve and Karhunen integral equation, the integrated mean square error is a minimum with respect to variations in the eigenfunctions. Furthermore, this minimum error is equal to the sum of the omitted eigenvalues. By means of this expansion, the prediction problem is reduced to that of determining a finite number of coefficients which constitute a set of independent random variables. These coefficients are specified by applying a minimum mean square criterion to a suitable non-linear or linear function of the known parts of the sample function to be extrapolated.

The method is also described in terms of discrete or sampled processes and is found to depend on an interesting theorem in matrix algebra.

Finally, the method is applied to the problem of estimating the electricity demand in a large area several hours in advance and some results are given.

## Sommaire

Le rapport décrit une méthode de prédiction d'un processus non stationnaire. Il montre que si la fonction échantillonnée caractérisant le processus est exprimée dans les termes de la fonction propre de l'équation intégrale de Loeve-Karhunen, le carré moyen de l'erreur est minimum par rapport aux variations de la fonction propre. L'erreur minimale est égale à la somme des valeurs propres omises. Ce développement réduit le problème de prédiction à la détermination d'un nombre fini de coefficients qui constituent un ensemble de variables aléatoires indépendantes; ces coefficients sont déterminés par l'application du critère du carré moyen minimum à une fonction appropriée, linéaire ou non, des termes connus de la fonction échantillonnée qui est à extrapoler.

La méthode est décrite dans le cas d'un processus échantillonné; il ressort qu'elle est basée sur un théorème très intéressant de l'algèbre matricielle.

Pour finir, cette méthode est appliquée au problème de la prévision plusieurs heures à l'avance de la consommation d'un grand réseau électrique et quelques résultats de l'application de cette méthode sont donnés.

## Zusammenfassung

Der Aufsatz beschreibt ein Verfahren zur Vorhersage eines nicht-stationären Prozesses und zeigt: Werden die Abtastfunktionen eines solchen Prozesses als Eigenfunktionen der Loeve-Karhunen'schen Integralgleichung entwickelt, so ist der quadratische Mittelwert des Fehlers in bezug auf Änderungen der Eigenfunktionen ein Minimum. Außerdem ist dieser minimale Fehler gleich der Summe der unterdrückten Eigenwerte. Diese Entwicklung reduziert das Vorhersageproblem auf die Bestimmung einer endlichen Anzahl von Koeffizienten, die eine Folge unabhängiger statistischer Variablen bilden. Diese Koeffizienten lassen sich näher bestimmen, wenn man ein Kriterium des kleinsten quadratischen Mittelwertes auf eine geeignete nichtlineare oder lineare Funktion der bekannten Teile der getasteten Funktion zur Extrapolation anwendet.

Bei der durchgeführten Erweiterung der Methode auf diskrete oder getastete Prozesse stellt sich heraus, daß sie auf einem interessanten Theorem der Matrizenrechnung beruht.

Schließlich wird die Methode dazu benützt, mehrere Stunden im voraus den Elektrizitätsbedarf in einem großen Gebiet abzuschätzen; einige Ergebnisse sind dargestellt.

## Introduction

The theory of Wiener<sup>1</sup>, in its original form, is concerned with the prediction and smoothing of stationary processes by linear filters; Wiener's method has, however, been extended by Booton<sup>2</sup> to the processing of non-stationary time-series by time-variant devices and by Zadeh<sup>3</sup> to non-linear filtering. However, Wiener's theory and its extensions do not take account of any constraints to which the process may be subject. For example, the process might contain natural 'modes', a knowledge of which would certainly facilitate its prediction.

In the following, it is shown that the characteristic modes of a process may be defined in an unambiguous and optimum manner. Methods of predicting the process from a knowledge of its modes are described.

## The Generalized Wiener Predictor

In the generalized form, the prediction problem is that of synthesizing a time-variant linear filter whose output,  $y(t)$  at any time  $t$ , is an estimate of the input at the later time  $t + \alpha$ . More precisely, if  $x(t)$  is a typical member of a non-stationary statistical process, the filter is required to minimize the mean square error  $[e^2(t)]$  between the output  $y(t)$  and the input  $x(t + \alpha)$  at time  $t + \alpha$ . If  $\varepsilon(t)$  is the error at time  $t$ , then

$$\varepsilon(t) = y(t) - x(t + \alpha)$$

The mean square error  $[e^2(t)]$  is obtained by squaring  $\varepsilon(t)$  and averaging over the group of inputs and outputs

$$[e^2(t)] = [y^2(t)] - [2y(t)x(t + \alpha)] + [x^2(t + \alpha)] \quad (1)$$

(Here the square brackets denote group averaging.)

Let  $g(t, \tau)$  denote the output of the filter at time  $t$  due to a unit impulse input (delta function) at time  $\tau$ ; then the output  $y(t)$  due to the input  $x(t)$  is given by

$$y(t) = \int_0^t g(t, \tau) x(\tau) d\tau \quad (2)$$

Combining eqns (1) and (2) gives

$$[e^2(t)] = \int_0^t \int_0^t g(t, \tau) g(t, \tau') R(\tau, \tau') d\tau d\tau' - 2 \int_0^t g(t, \tau) R(t + \alpha, \tau) d\tau + R(t + \alpha, t + \alpha) \quad (3)$$

where  $R(\tau, \tau')$  denotes the autocorrelation function of the process

$$R(\tau, \tau') = [x(\tau)x(\tau')] \quad (4)$$

If the response function  $g(t, \tau)$  is varied by  $\delta g(t, \tau)$  then (to a first order) the corresponding variation in  $[\varepsilon^2(t)]$  is

$$[\delta \varepsilon^2(t)] = 2 \int_0^t \delta g(t, \tau) \left\{ \int_0^t R(\tau, \tau') g(t, \tau') d\tau' - R(t+\alpha, \tau) \right\} d\tau$$

If  $g(t, \tau)$  is the response function of the optimum filter which minimizes  $[\varepsilon^2(t)]$ , then the variation  $[\delta \varepsilon^2(t)]$  vanishes for all  $\delta g(t, \tau)$  and hence

$$\int_0^t R(\tau, \tau') g(t, \tau') d\tau' = R(t+\alpha, \tau) \quad 0 < \tau < t \quad (5)$$

The solution  $g(t, \tau)$  of this modified Wiener-Hopf integral equation is the response of the optimum Wiener filter and the resulting mean square error is given by

$$[\varepsilon^2(t)]_{\min} = R(t+\alpha, t+\alpha) - \int_0^t g(t, \tau) R(t+\alpha, \tau) d\tau \quad (6)$$

The first term on the right is the mean power  $[x^2(t+\alpha)]$ , flowing at the time  $t+\alpha$ , and the second term represents the maximum difference between the mean power and the mean square error that it is possible to obtain by linear filtering.

The response function  $g(t, \tau)$  of the Wiener filter is specified uniquely [cf. eqn (5)] by the autocorrelation function  $R(\tau, \tau')$  of the process. The applicability of the predictor depends, therefore, on the possibility of constructing a representative and realistic autocorrelation function which, in practice, must be derived from a finite number of the sample functions of the process. Suppose, for example, that the autocorrelation function  $R(\tau, \tau')$  is formed from the  $M$  sample functions  $x_m(t)$ , ( $m = 1, 2, 3, \dots, M$ ) then

$$R(\tau, \tau') = \frac{1}{M} \sum_{m=1}^M x_m(\tau) x_m(\tau')$$

The sample functions  $x_m(t)$  are determined ideally by performing  $M$  sets of measurement under identical conditions. However, in many practical situations the number of sample functions it is possible to measure under almost identical conditions is severely limited by uncontrollable variations in one or more of the parameters of the process. For example, if the functions  $x_m(t)$  were daily curves of electricity demand, then the number which it is practically useful to include, in forming the autocorrelation function, is limited by the seasonal trend of the load and by the economic growth in demand. However, for a large class of stochastic processes, the sample functions are expressible in terms of a small number of 'characteristic modes' which arise from the physical nature of the source of the particular process. Under such conditions the inclusion of a large number of sample functions in calculating the autocorrelation function is neither necessary nor desirable. Furthermore it is the 'modes' rather than the correlation function alone which characterizes the process and the Wiener predictor is no longer appropriate.

#### Derivation of the Characteristic Modes

Suppose that  $x_m(t)$ ,  $m = 1, 2, 3, \dots, M$ , are  $M$  sample functions of a stochastic process and that it is required to define, in

some sense, the characteristic modes of the process. A simple way of specifying the first mode is to seek a function  $\phi_1(t)$ , a scaling factor  $\lambda_1^{\frac{1}{2}}$  and a set of coefficients  $a_{m1}$  such that  $\lambda_1^{\frac{1}{2}} a_{m1} \phi_1(t)$  approximates, in a least-squares sense, to the sample functions  $x_m(t)$  over a specified time interval  $(0, T)$ . The  $m$ th error takes the form

$$\varepsilon_m(t) = x_m(t) - \lambda_1^{\frac{1}{2}} a_{m1} \phi_1(t)$$

The mean square error averaged over time and over the group is:

$$E = \frac{1}{MT} \sum_{m=1}^M \int_0^T \{x_m(t) - \lambda_1^{\frac{1}{2}} a_{m1} \phi_1(t)\}^2 dt \quad (7)$$

The constants  $\lambda_1^{\frac{1}{2}} a_{m1}$  and the function  $\phi_1(t)$  may be chosen such that this error  $E$  is a minimum. This minimum is attained when the variation in  $E$  due to first order variations in  $\lambda_1^{\frac{1}{2}} a_{m1}$  and  $\phi_1(t)$  vanishes, i.e. when

$$\begin{aligned} \int_0^T \phi_1(t) x_m(t) dt &= \lambda_1^{\frac{1}{2}} a_{m1} \int_0^T \phi_1^2(t) dt \quad m = 1, 2, \dots, M \\ \sum_{m=1}^M a_{m1} x_m(t) &= \sum_{m=1}^M \lambda_1^{\frac{1}{2}} a_{m1}^2 \phi_1(t) \quad 0 < t < T \end{aligned}$$

The function  $\phi_1(t)$  and the vector  $a_{m1}$  may, without loss of generality, be normalized such that

$$\begin{aligned} \int_0^T \phi_1^2(t) dt &= 1 \\ \sum_{m=1}^M a_{m1}^2 &= M \end{aligned}$$

The minimum conditions then take the form

$$\begin{aligned} a_{m1} &= \lambda_1^{-\frac{1}{2}} \int_0^T \phi_1(t) x_m(t) dt \\ \lambda_1^{\frac{1}{2}} \phi_1(t) &= \frac{1}{M} \sum_{m=1}^M a_{m1} x_m(t) \end{aligned} \quad (8)$$

Elimination of  $a_{m1}$  between these two relations yields

$$\frac{1}{M} \sum_{m=1}^M \int_0^T \phi_1(\tau) x_m(\tau) d\tau x_m(t) = \lambda_1 \phi_1(t)$$

or

$$\int_0^T R(t, \tau) \phi_1(\tau) d\tau = \lambda_1 \phi_1(t) \quad 0 < t < T \quad (9)$$

where  $R(t, \tau)$  is the correlation function formed from the group of the  $M$  sample functions

$$R(t, \tau) = \frac{1}{M} \sum_{m=1}^M x_m(t) x_m(\tau) \quad (10)$$

Combining eqns (7) and (9) gives for the minimum mean square error

$$E = \frac{1}{T} \left\{ \int_0^T R(\tau, \tau) d\tau - \lambda_1 \right\} \quad (11)$$

Eqn (9) shows that  $\phi_1(t)$  is an eigenfunction of the modified Wiener-Hopf integral equation. However, eqn (11) indicates that the error is a minimum when  $\lambda_1$  is the greatest or dominant eigenvalue and hence  $\phi_1(t)$  is the dominant eigenfunction. The

second mode may be, similarly, specified by requiring that  $\lambda_2^{\frac{1}{2}} a_{m2} \phi_2(t)$  be the best mean-square approximation to the error  $\varepsilon_m(t)$  and it follows that  $\lambda_2$  is the second largest eigenvalue and  $\phi_2(t)$  is the corresponding eigenfunction. Continuing in this way an expansion for  $x_m(t)$  is obtained and this takes the form

$$x_m(t) = \lambda_1^{\frac{1}{2}} a_{m1} \phi_1(t) + \lambda_2^{\frac{1}{2}} a_{m2} \phi_2(t) + \lambda_3^{\frac{1}{2}} a_{m3} \phi_3(t) + \dots$$

where  $\phi_k(t)$  and  $\lambda_k$  satisfy the eigenfunction equation

$$\int_0^T R(t, \tau) \phi_k(\tau) d\tau = \lambda_k \phi_k(t) \quad 0 < t < T \quad k = 1, 2, 3, \dots \quad (12)$$

and where the coefficients  $a_{mk}$  are given by

$$a_{mk} = \lambda_k^{-\frac{1}{2}} \int_0^T \phi_k(t) x_m(t) dt \quad (13)$$

It follows from the general theory<sup>4</sup> of the eigenvalue equation, that the functions  $\phi_k(t)$  are orthogonal and as they are normalized form an orthonormal set

$$\int_0^T \phi_k(\tau) \phi_{k'}(\tau) d\tau = \delta_{kk'} = \begin{cases} 1 & \text{if } k' = k \\ 0 & \text{if } k' \neq k \end{cases} \quad (14)$$

The constant  $a_{mk}$  may be regarded as the value which a random variable  $a_k$  takes for the  $m$ th number of the group. The correlation between the random variables  $a_k$  and  $a_{k'}$  then takes the form

$$[a_k a_{k'}] = \frac{1}{M} \sum_{m=1}^M a_{mk} a_{mk'}$$

From eqn (13) it then follows that

$$\begin{aligned} [a_k a_{k'}] &= \frac{1}{M} \sum_{m=1}^M (\lambda_k \lambda_{k'})^{-\frac{1}{2}} \int_0^T \int_0^T x_m(\tau) x_m(\tau') \phi_k(\tau) \phi_{k'}(\tau') d\tau d\tau' \\ &= (\lambda_k \lambda_{k'})^{-\frac{1}{2}} \int_0^T \int_0^T R(\tau, \tau') \phi_k(\tau) \phi_{k'}(\tau') d\tau d\tau' \\ &= (\lambda_k \lambda_{k'})^{-\frac{1}{2}} \lambda_k \int_0^T \phi_k(\tau) \phi_{k'}(\tau) d\tau \\ &= \delta_{kk'} \quad \text{i.e. } [a_k a_{k'}] = \delta_{kk'} \end{aligned} \quad (15)$$

Thus the random variables  $a_k$  possess unit mean square values and zero cross correlations. It is evident that the mode functions  $\phi_k(\tau)$  are identical to those introduced by Loève<sup>5</sup> and Karhunen<sup>6</sup> in deriving an expansion with uncorrelated coefficients; for this reason eqn (12) might be termed the Loève-Karhunen integral equation.

The autocorrelation function  $R(\tau, \tau')$  may also be expanded in terms of the characteristic modes of the process

$$R(\tau, \tau') = \sum_{k=1}^{\infty} \lambda_k \phi_k(\tau) \phi_k(\tau') \quad (16)$$

It may also be shown that all the eigenvalues  $\lambda_k$  are either positive or zero

$$\lambda_k > 0 \quad (17)$$

An important property of the expansion of  $x_m(t)$  is that, if the series is terminated after  $K$  terms, the time integrated mean square error  $ET$  takes the form

$$ET = \int_0^T R(\tau, \tau') d\tau' - \sum_{k=1}^K \lambda_k \quad (18)$$

The term

$$\int_0^T R(\tau, \tau') d\tau$$

is equal to the average energy contained in the process over the time interval  $(0, T)$ . By virtue of eqns (16) and (14) this energy takes the form

$$\int_0^T R(\tau, \tau) d\tau = \sum_{k=1}^{\infty} \lambda_k \quad (19)$$

Combining eqns (18) and (19) gives

$$ET = \sum_{k=K+1}^{\infty} \lambda_k \quad (20)$$

Thus the integrated square error  $ET$  is equal to the sum of the omitted eigenvalues and the number of terms required for a specified accuracy is readily obtainable from eqn (18) or (20). The eigenvalue  $\lambda_k$ , being essentially non-negative, may be interpreted as the energy associated with the  $k$ th mode of the process.

The impulse response of the Wiener predictor may be described in terms of the modes  $\phi_k(t)$ . Suppose, for example, that the predictor is required to estimate an input sample function at the time  $T = t + \alpha$  from a knowledge of the input over the time interval  $(0, T_0)$  where  $T_0 < T$ . The impulse response satisfies the relation

$$\int_0^{T_0} R(\tau, \tau') g(T_0, \tau') d\tau' = R(T, \tau) \quad 0 < \tau < T_0$$

If  $R(\tau, \tau')$  is expanded in the form of eqn (16) the integral equation takes the form

$$\sum_k \lambda_k g_k \phi_k(\tau) = \sum_k \lambda_k \phi_k(T) \phi_k(\tau) \quad 0 < \tau < T_0$$

where

$$g_k = \int_0^{T_0} g(T_0, \tau) \phi_k(\tau) d\tau \quad (21)$$

Multiplication by  $\phi_{k'}(\tau)$  and integration over the range  $(0, T_0)$  yields

$$\sum_{k'} A_{kk'} \lambda_{k'} g_{k'} = \sum_{k'} A_{kk'} \lambda_{k'} \phi_{k'}(T) \quad (22)$$

where

$$A_{kk'} = \int_0^{T_0} \phi_k(\tau) \phi_{k'}(\tau) d\tau \quad (23)$$

As the function given by the series

$$\sum_k g_k \phi_k(\tau)$$

must vanish in the range  $T_0 < \tau < T$  it follows that

$$\sum_{k'} B_{kk'} g_{k'} = 0 \quad (24)$$

where

$$B_{kk'} = \int_{T_0}^T \phi_k(\tau) \phi_{k'}(\tau) d\tau \quad (25)$$

The matrices  $A$  and  $B$  with elements  $A_{kk'}$  and  $B_{kk'}$  respectively are idem-potent in that they satisfy the relations

$$A^2 = A, \quad B^2 = B \quad (26)$$

Furthermore, as the mode functions are orthonormal over the whole range  $(0, T)$  it follows that

$$\begin{aligned} A + B &= I \\ AB &= BA = 0 \end{aligned} \quad (27)$$

where  $I$  denotes the unit matrix. In terms of the mode functions the mean square prediction error of eqn (3) takes the form

$$[\varepsilon^2(T_0)] = \sum_{k=1}^{\infty} \lambda_k \{g_k - \phi_k(T)\}^2 \quad (28)$$

This error is minimized when  $g_k$  satisfies eqns (22) and (24), i.e. when

$$\begin{aligned} \sum_{k'} A_{kk'} \lambda_{k'} g_{k'} &= \sum_{k'} A_{kk'} \lambda_{k'} \phi_{k'}(T) \\ \sum_{k'} B_{kk'} g_{k'} &= 0 \end{aligned} \quad (29)$$

It follows from eqns (26) and (27) that the sum of the ranks of the two matrices  $A$  and  $B$  is equal to the rank of the unit matrix  $I$ . Consequently in eqn (29) there are exactly as many independent equations as there are unknowns  $g_k$  and the solution is unique. These simultaneous equations are completely equivalent to the Wiener-Hopf relation (5).

The contribution of the term in  $g_k$  to the prediction error is proportional to the energy or eigenvalue  $\lambda_k$  [this is demonstrated by eqn (28)]. It follows then that any error in the value of  $g_k$  does not appreciably increase the prediction error provided that the corresponding eigenvalue  $\lambda_k$  is sufficiently small. In particular, those coefficients  $g_k$  corresponding to negligible eigenvalues  $\lambda_k$  may be neglected, i.e. modes with negligible mean energies may be ignored.

### Prediction of a Process by its Characteristic Modes

The prediction problem is essentially that of estimating the values of a sample function  $x(t)$  of a non-stationary process in the time interval  $(T_0, T)$  from a knowledge of  $x(t)$  in the range  $(0, T_0)$ .

It has been established, eqn (18), that if the process is expanded to  $K$  terms as a combination of its characteristic modes then the integrated mean square error  $ET$  takes the form

$$ET = \int_0^T R(\tau, \tau) d\tau' - \sum_{k=1}^K \lambda_k$$

From this criterion the number  $K$  of terms which achieve a specified accuracy may be derived. To this order accuracy the sample function  $x(t)$  may be expressed in the form

$$x(t) = \sum_{k=1}^K c_k \phi_k(t) \quad 0 < t < T \quad (30)$$

where  $c_k$  are a set of constants. This expansion is valid over the whole time interval  $(0, T)$  including the part of the range in which  $x(t)$  is to be predicted, provided that the coefficients  $c_k$  are specified appropriately. The prediction problem, therefore, reduces to that of determining the set of constants  $c_k$ . This

method of prediction automatically treats the sample function  $x(t)$  as a combination of its characteristic modes whereas the Wiener predictor imposes no such constraints on  $x(t)$  treating it quite generally. There are several methods of specifying the coefficients  $c_k$  depending on the error criterion adopted. Two possible methods are described below.

*Method 1*—A particularly simple procedure is to choose the coefficients  $c_k$  such that eqn (30) is the best mean square approximation to  $x(t)$  over the time interval  $(0, T_0)$  for which  $x(t)$  is known.

If  $\varepsilon(t)$  denotes the expansion error then

$$\varepsilon(t) = x(t) - \sum_{k=1}^K c_k \phi_k(t)$$

The integrated square error  $D$  is given by

$$D = \int_0^{T_0} \varepsilon^2(t) dt$$

and hence

$$D = \int_0^{T_0} \left\{ x(t) - \sum_{k=1}^K c_k \phi_k(t) \right\}^2 dt \quad (31)$$

The error  $D$  is a minimum when

$$\frac{\partial D}{\partial c_k} = 0 \quad k = 1, 2, 3, \dots, K$$

and hence when

$$\sum_{k'=1}^K c_{k'} \int_0^{T_0} \phi_k(\tau) \phi_{k'}(\tau) d\tau = \int_0^{T_0} x(\tau) \phi_k(\tau) d\tau \quad k = 1, 2, \dots, K \quad (32)$$

This set of simultaneous linear equations determines the coefficients  $c_k$  uniquely. The prediction of  $x(t)$  is effected by applying eqn (30) over the time interval  $(T_0, T)$ .

*Method 2*—An alternative method of specifying the coefficients  $c_k$  is to require that they approximate to the true coefficients in a least-squares sense. The true coefficients, as is evident from eqn (8), take the form  $d_k$  where

$$d_k = \int_0^T x(t) \phi_k(t) dt \quad (33)$$

On the other hand, the constants  $c_k$  depend only on the values of  $x(t)$  in the range  $(0, T_0)$  and it has been shown by Volterra<sup>7</sup> that the most general operation of this type on a function  $x(t)$  must take the form

$$c_k = \int_0^{T_0} f_k(\tau) x(\tau) d\tau + \int_0^{T_0} \int_0^{T_0} h_k(\tau, \tau') x(\tau) x(\tau') d\tau d\tau' + \dots \quad (34)$$

If  $\varepsilon_k$  denotes the error between  $c_k$  and  $d_k$  then, using eqn (33), the following relation is obtained for the mean square error  $[\varepsilon_k^2]$

$$\begin{aligned} [\varepsilon_k^2] &= \int_0^T \int_0^T R(\tau, \tau') \phi_k(\tau) \phi_k(\tau') d\tau d\tau' \\ &\quad - 2 \int_0^T [c_k x(\tau)] \phi_k(\tau) d\tau + [c_k^2] \end{aligned} \quad (35)$$

The functions  $f_k(\tau)$ ,  $h_k(x, \tau')$ , etc. which occur in eqn (34) may be chosen so as to minimize  $[\varepsilon_k^2]$ . To effect this minimization  $c_k$  is substituted for in eqn (35) and gives

$$\text{Eqn (36)}^*$$

The error  $[\varepsilon_k^2]$  is a minimum when its first variation with respect to variations in  $f_k(\tau)$ ,  $h_k(\tau', \tau)$ , etc. vanishes, i.e. when

$$\text{Eqn (37)}^\dagger$$

The functions  $f_k(\tau)$ ,  $h_k(\tau, \tau')$  etc. which satisfy eqn (37) define an optimum non-linear predictor for the process  $x(t)$ . The determination of this predictor requires, however, a knowledge of the higher order covariances  $[x(\tau)x(\tau')x(\tau'')]$  etc. If this treatment is restricted here to the case of linear predictors the functions  $h_k(\tau, \tau')$  etc. may be neglected and eqn (37) takes the comparatively simple form

$$\int_0^{T_0} R(\tau, \tau') f_k(\tau') d\tau' = \lambda_k \phi_k(\tau) \quad 0 < \tau < T_0 \quad (38)$$

For each value of  $k$  in the range  $1 \leq k \leq K$ , eqn (38) defines a function  $f_k(\tau)$  which, in turn, determines the weighting coefficient  $c_k$  of the  $k$ th mode according to the relation

$$c_k = \int_0^{T_0} x(\tau) f_k(\tau) d\tau \quad (39)$$

The predicted values of  $x(t)$  are obtained from eqn (30) as before. A matrix formulation of the prediction method which lends itself directly to digital computation is described in the Appendix.

#### Application to the Problem of Load Estimation

The problem considered here is that of predicting the electricity demand in a large area, up to several hours ahead. Predictions of load are required for the purposes of ordering generating plant, loading of plant and checking the security of the supply network in advance. If automatic control is to be applied to a supply system then the prediction of load by digital computation becomes a necessity.

For the purpose of prediction, each daily load curve may be divided into part-day periods of several hours' duration, the

periods including the interval  $(T_0, T)$  over which the prediction is required together with an immediately preceding interval  $(0, T_0)$ ; the period as a whole is then coextensive with the period  $(0, T)$ . The load during each of these periods depends on many factors including meteorological parameters and the response of the consumer to television and radio programmes, etc.; of these factors the meteorological ones have the greatest influence. It is plausible, therefore, to express the load  $x_{mn}$  on the  $m$ th period at the  $n$ th instant of time in the form

$$x_{mn} = \alpha_n + f_1(T_m) \beta_n + f_2(L_m) \gamma_n + f_3(W_m) \delta_n + \dots \quad (40)$$

where  $f_1(T_m)$ ,  $f_2(L_m)$ ,  $f_3(W_m)$ , are functions of temperature  $T_m$ , light intensity  $L_m$ , wind velocity  $W_m$  respectively. The quantity  $\alpha_n$  represents the base load and the factors  $\beta_n$ ,  $\gamma_n$ ,  $\delta_n$ , allow for the varying importance of weather parameters with time of day. According to eqn (40) each load vector is linearly dependent on the vectors  $\alpha$ ,  $\beta$ ,  $\gamma$ ,  $\delta$ , ..., however, if the load is expanded to  $K$  terms as a combination of its characteristics modes, the quantity  $x_{mn}$  takes the form

$$x_{mn} = \sum_{k=1}^K c_{mk} \phi_{kn} \quad (41)$$

It has been shown previously that the mode vectors  $\phi_k$  minimize the error of the expansion and it is evident that if the mode vectors were replaced by  $K$  linearly independent combinations then the error is unchanged. Thus the  $K$  dimensional manifold formed by the mode vectors and their linear combinations gives a smaller error than any other  $K$  dimensional manifold. It is plausible, therefore, to regard the vectors  $\alpha$ ,  $\beta$ ,  $\gamma$ ,  $\delta$ , ... of eqn (40) as belonging to the manifold of mode vectors. It is concluded that the modes describe the basic trends of the load under average weather conditions for the period of the records whereas the weighting coefficients  $c_{mk}$  depend on the meteorological parameters relevant to the  $m$ th period.

#### Results

The method of characteristic modes has been used to predict loads up to 8 h ahead in a large area with a peak demand of some 5,000 MW. The weighting coefficients were calculated by the procedure described as Method 1. Computer results,

\* Eqn (36):

$$\begin{aligned} [\varepsilon_k^2] = & \lambda_k - 2 \int_0^{T_0} \lambda_k f_k(\tau) \phi_k(\tau) d\tau + \int_0^{T_0} \int_0^{T_0} R(\tau, \tau') f_k(\tau) f_k(\tau') d\tau d\tau' \\ & + 2 \int_0^{T_0} \int_0^{T_0} \int_0^{T_0} [x(\tau)x(\tau')x(\tau'')] f_k(\tau) h_k(\tau', \tau'') d\tau d\tau' d\tau'' - 2 \int_0^{T_0} d\tau \int_0^{T_0} d\tau' \int_0^{T_0} d\tau'' [x(\tau)x(\tau')x(\tau'')] h_k(\tau', \tau'') \phi_k(\tau) \\ & + \int_0^{T_0} \int_0^{T_0} \int_0^{T_0} \int_0^{T_0} [x(\tau)x(\tau')x(\tau'')x(\tau''')] h_k(\tau, \tau') h_k(\tau'', \tau''') d\tau d\tau' d\tau'' d\tau''' + \dots \end{aligned} \quad (36)$$

† Eqn (37):

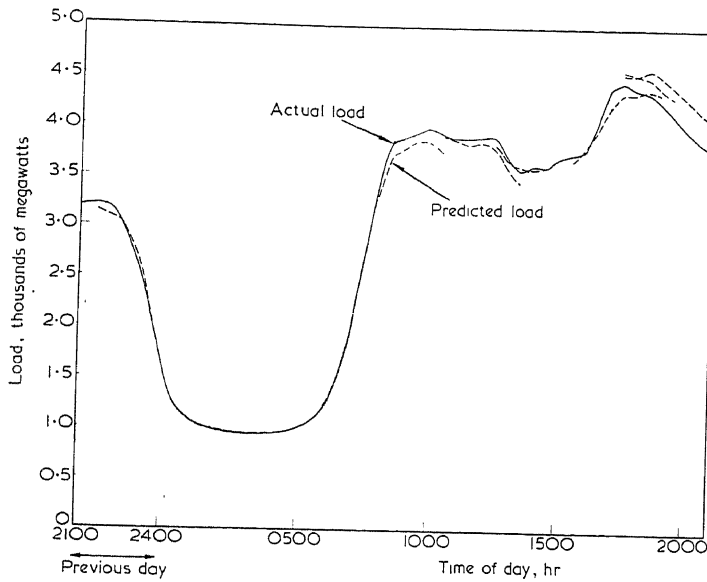
$$\begin{aligned} & \int_0^{T_0} R(\tau, \tau') f_k(\tau') d\tau' - \lambda_k \phi_k(\tau) + \int_0^{T_0} \int_0^{T_0} [x(\tau)x(\tau')x(\tau'')] h_k(\tau', \tau'') d\tau d\tau'' + \dots = 0 \\ & \int_0^{T_0} [x(\tau)x(\tau')x(\tau'')] f_k(\tau) d\tau - \int_0^T [x(\tau)x(\tau')x(\tau'')] \phi_k(\tau) d\tau + \int_0^{T_0} \int_0^{T_0} [x(\tau)x(\tau')x(\tau'')] h_k(\tau, \tau'') d\tau d\tau'' + \dots = 0 \quad (37) \\ & \text{etc.} \end{aligned}$$

obtained to date, give predictions for 30 weekdays covering six different periods each day. The characteristic modes were calculated, using the data of up to 20 immediately preceding weekdays. The calculations were performed on an IBM 7090 computer and about 1 min of machine time was used in the calculation of each predicted curve. It was found that use of five or six eigenvectors gave the best results, depending on the length of the prediction and that the R.M.S. error was about 3 per cent for predictions of 3 to 4 h ahead.

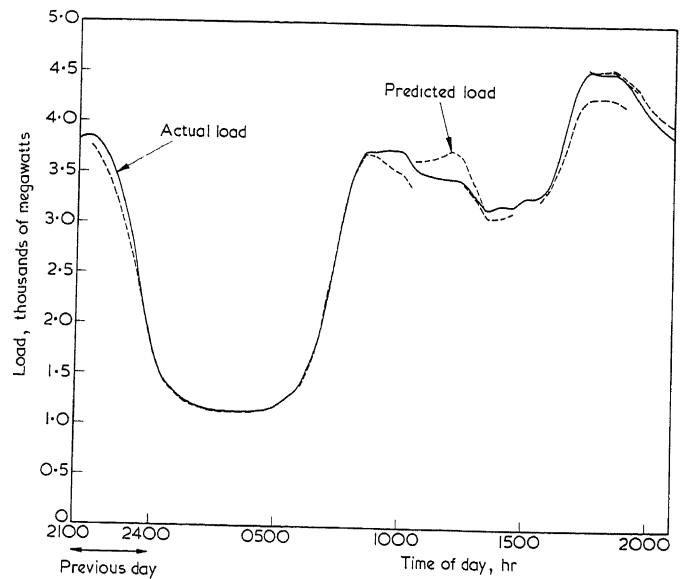
Curves comparing predicted load with actual load are illustrated in *Figures 1-8* for eight typical weekdays. Each of predicted curves was calculated 30 min before its commence-

ment. The most important predictions, for the purpose of ordering generating plant in advance, are those which cover the breakfast peak at about 08.30 hours and the evening peak at about 18.00 hours. Results obtained for the breakfast peak compare favourably with predictions made at control centres using meteorological data, whereas the evening predictions were found to be less accurate on average.

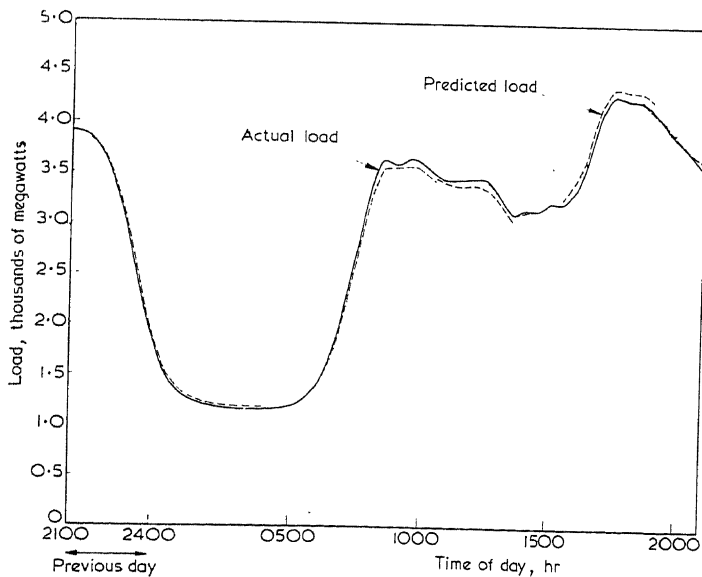
The problem of estimating the load from its characteristic modes is being further investigated. In particular the procedure described as Method 2 will be tested although no results are available at the present time. In addition the problem of deriving confidence limits for the predictions is being studied.



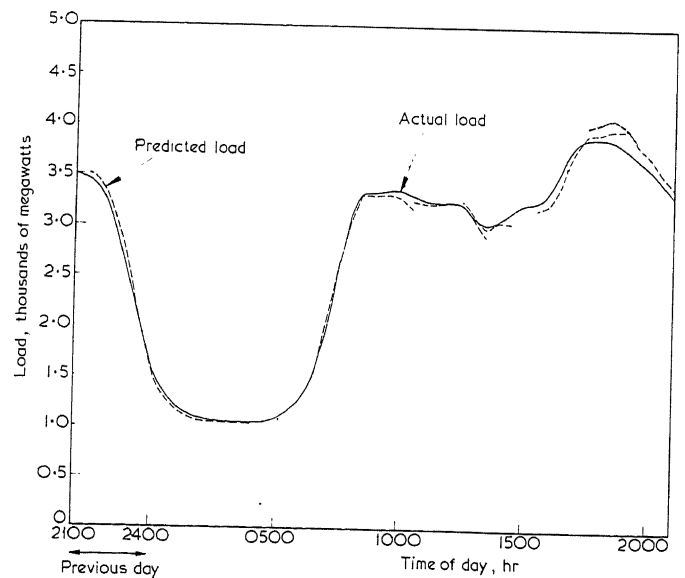
*Figure 1. Predicted and actual load on Monday 27th November*



*Figure 2. Predicted and actual load on Tuesday 28th November*



*Figure 3. Predicted and actual load on Wednesday 29th November*



*Figure 4. Predicted and actual load on Friday 1st December*

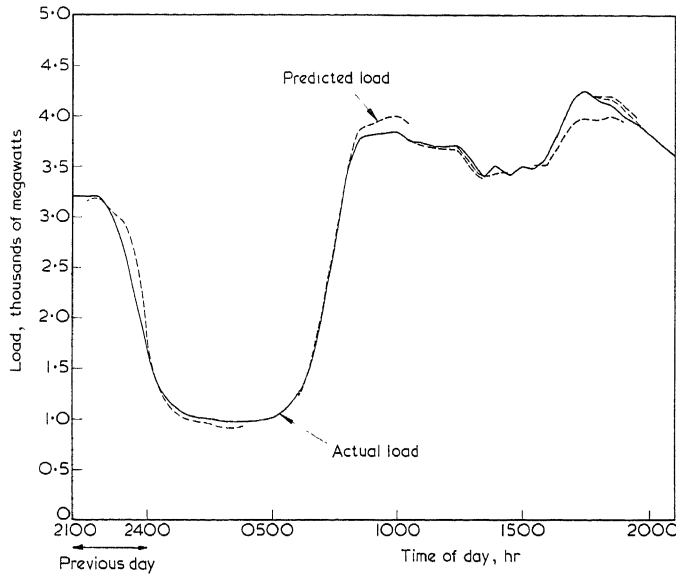


Figure 5. Predicted and actual load on Monday 4th December

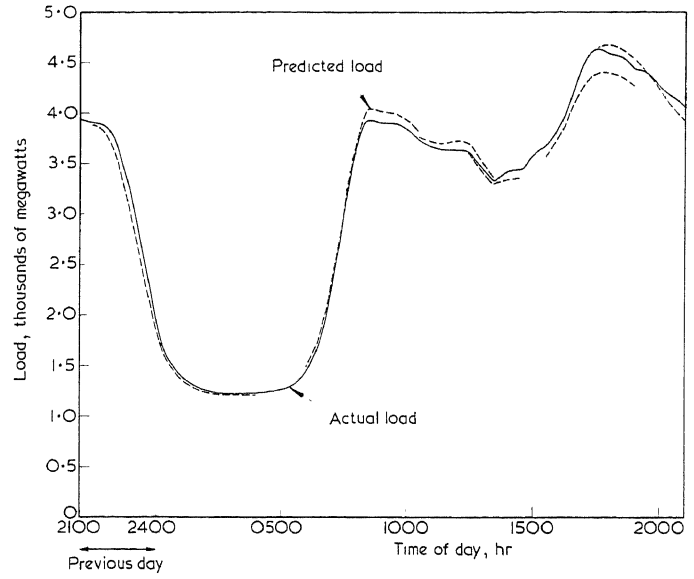


Figure 6. Predicted and actual load on Wednesday 6th December

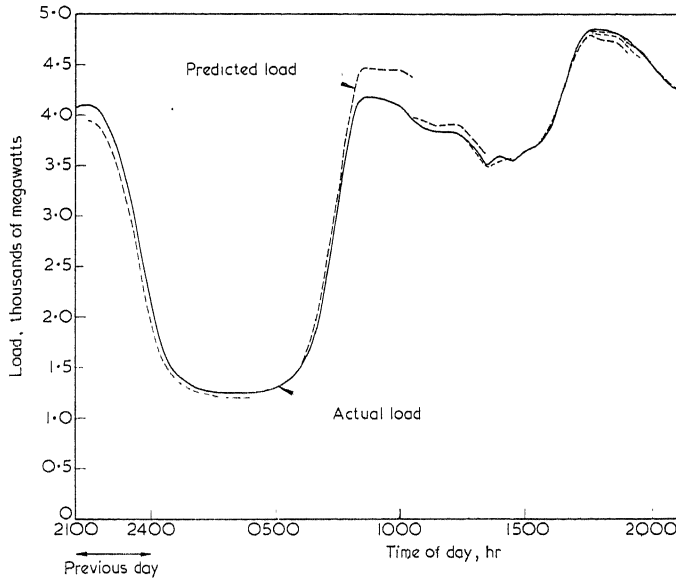


Figure 7. Predicted and actual load on Thursday 7th December

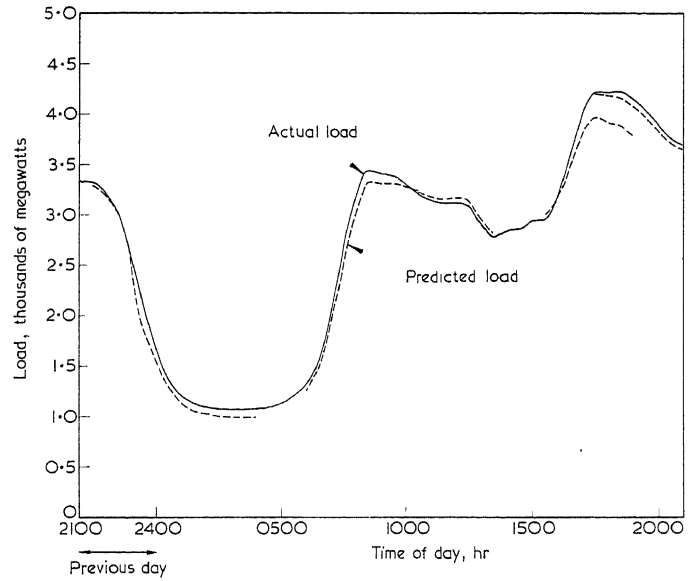


Figure 8. Predicted and actual load on Thursday 14th December

## Appendix

### Matrix Formulation

For the purposes of digital computation it is more convenient to treat the process  $x(t)$  as a discrete function of time rather than a continuous one. Suppose, then, that the process is sampled at instants of time separated by equal intervals, the sampling frequency being sufficiently high compared with the reciprocal bandwidth, and let  $x_{mn}$  denote the value of the  $m$ th member of the group at  $n$ th instant of time. The elements  $x_{mn}$  form a rectangular matrix  $X$ ,

$$X = (x_{mn})$$

If there are  $M$  members of the group and  $N$  sampling instants then  $X$  has  $M$  rows and  $N$  columns. Corresponding to the auto-

correlation  $R(\tau, \tau')$  there exists a covariance matrix  $R$  which is defined by

$$R = \frac{1}{M} X' X \quad (42)$$

where  $X'$  denotes the transpose of the matrix  $X$ .

The expansion of the process in terms of its characteristic modes is equivalent to expressing the elements of the matrix  $X$  in the form

$$x_{mn} = \sum_{k=1}^K a_{mk} \lambda_k^{\frac{1}{2}} \phi_{kn} \quad (43)$$

Denoting the errors by  $\varepsilon_{mn}$ , i.e. writing

$$\varepsilon_{mn} = x_{mn} - \sum_{k=1}^K a_{mk} \lambda_k^{\frac{1}{2}} \phi_{kn}$$

and choosing  $a_{mk}$  and  $\phi_{kn}$  such that the sum of the squares of  $\varepsilon_{mn}$  is a minimum gives

$$\begin{aligned}\lambda_k^{\frac{1}{2}} a_{mk} &= \sum_{n=1}^N x_{mn} \phi_{kn} \\ M \lambda_k^{\frac{1}{2}} \phi_{kn} &= \sum_{m=1}^M a_{mk} x_{mn}\end{aligned}\quad (44)$$

This set of equations is equivalent to eqn (8). Letting  $P$  denote the matrix with elements  $a_{mk}$ ,  $Q$  the matrix with elements  $\phi_{kn}$  and  $A$  the diagonal matrix with elements  $\lambda_k$  then eqn (43) takes the form

$$X = P A^{\frac{1}{2}} Q \quad (45)$$

and eqns (44) become

$$\begin{aligned}P A^{\frac{1}{2}} &= X Q' \\ M A^{\frac{1}{2}} Q &= P' X\end{aligned}\quad (46)$$

Eliminating  $P$  and  $Q$  in turn from eqn (46) gives the result

$$\begin{aligned}R Q' &= A Q' \\ S P &= P A\end{aligned}\quad (47)$$

where  $S$  is the matrix given by

$$S = \frac{1}{M} X X'$$

The first relation of (47) is the matrix form of the integral eqn (12).

It has been established that for any rectangular matrix  $X$  the best approximation of the form (45) is attained when  $P$  and  $Q$  satisfy the eigenvalue eqns (47).

In its matrix form, the prediction problem is that of estimating the values of a vector  $x_n$  in the range  $n = N' + 1, n' + 2, \dots, N$  given the values of  $x_n$  in the range  $n = 1, 2, 3, \dots, N'$ . Corresponding to eqn (30),  $x_n$  is expressed in the form

$$x_n = \sum_{k=1}^K c_k \phi_{kn} \quad (48)$$

where  $\phi_{kn}$  are the elements of  $Q$  which is specified by eqn (47). Method 1 determines the coefficients  $c_k$  according to the relations

$$\sum_{k'=1}^K c_{k'} \left\{ \sum_{n=1}^N \phi_{kn} \phi_{k'n} \right\} = \sum_{n=1}^{N'} \phi_{kn} x_n \quad k=1, 2, \dots, K \quad (49)$$

These are  $K$  simultaneous equations for the  $K$  unknowns  $c_k$ . In the linear form of Method 2 the coefficients  $c_k$  are written in the form

$$c_k = \sum_{n=1}^N x_n f_{kn} \quad k=1, 2, \dots, K \quad (50)$$

Corresponding to eqn (50) the constants  $f_{kn}$  are specified by the relations

$$\sum_{n=1}^{N'} R_{nn'} f_{kn'} = \lambda_k \phi_{kn} \quad \begin{matrix} n=1, 2, 3, \dots, N' \\ k=1, 2, 3, \dots, K \end{matrix} \quad (51)$$

The sum of eigenvalues of the matrix  $R$  is equal to the sum of its diagonal elements or its trace; the error criterion analogous to eqn (18), therefore takes the form

$$NE = \text{trace } R - \sum_{k=1}^K \lambda_k = \sum_{k=K+1}^{\infty} \lambda_k \quad (52)$$

*The author wishes to thank the Director of the Central Electricity Research Laboratory for permission to publish this paper.*

## References

- <sup>1</sup> WIENER, N. *The Extrapolation, Interpolation and Smoothing of Stationary Time-Series*. 1949. New York; Wiley
- <sup>2</sup> BOOTON, B. C. An optimisation theory for time-varying linear systems with non-stationary statistical inputs. *Proc. Inst. Radio Engrs* 38 (1952)
- <sup>3</sup> ZADEH, L. A. Optimum non-linear filters. *J. appl. Phys.* 24 (1953)
- <sup>4</sup> COURENT, and HILBERT *Methods of Mathematical Physics*. Vol. 1. 1953. New York; Interscience
- <sup>5</sup> LOEVE, M. *C. R. Acad. Sci., Paris*. 220 (1946)
- <sup>6</sup> KARHUNEN, K. *Ann. Acad. Sci. fenn.* 37 (1947)
- <sup>7</sup> VOLTERRA, V. *Theory of Functionals*. 1930. London; Blackie



# On the Design of Predictor Control Systems

S. HORING

## Summary

The purpose of this paper is to investigate the optimum design of predictor control systems subject to statistical or deterministic inputs, under appropriate constraints on the control signal. The control signal constraints treated in this paper are: (a) the magnitude of the control signal equal to a constant (this is the bang-bang control problem), and (b) the magnitude of the control signal less than or equal to some constant.

The basic optimization problem is viewed, in very general terms, as a problem in decision theory. Based on available information, the decision must be as to which of the allowable control signals should be applied to the plant in order to achieve the optimum performance. A sum of squared error criterion is used to define the optimum in order to reduce the problem to a problem in linear decision theory. A piecewise-linear control boundary is found which can be used to generate the desired control signal. The realization of this control boundary is shown to be straightforward and inexpensive. This technique can be applied to linear time invariant plants of arbitrary order. Experimental results verifying the techniques which are developed are presented. Included in these is the control of the step response of a system with a saturation-limited unstable second order plant in the presence of noise.

## Sommaire

Le but de ce rapport est d'étudier le mode de réalisation optimale d'un système de prédiction soumis à des variations prédéterminées ou statistiques et compte tenu de certaines contraintes imposées au signal de commande. Les contraintes prises en considération dans ce rapport sont les suivantes: (a) l'amplitude du signal de commande est égale à une constante ('bang-bang control problem') et (b) l'amplitude de ce signal est égale ou inférieure à une valeur prescrite.

Le principe de ce problème d'optimisation est examiné tout d'abord sous sa forme la plus générale, comme un problème de la théorie des décisions. En utilisant les informations disponibles il s'agit de choisir la commande à appliquer au système en vue de réaliser une performance optimale; le critère de la somme des carrés moyens des erreurs est utilisé pour ramener ce problème à celui traité par la théorie de la décision linéaire. Une linéarisation des contraintes peut être trouvée permettant de produire le signal désiré. Le rapport indique que la réalisation d'une telle contrainte est efficace et relativement peu coûteuse. Cette technique peut être utilisée pour des installations linéaires invariantes dans le temps d'un ordre quelconque. Des résultats expérimentaux vérifiant le bien-fondé de cette technique, sont présente. Parmi ces résultats figurent le réglage de la réponse indicelle d'un système avec une installation instable du 2me ordre présentant des phénomènes de saturation en présence du bruit.

## Zusammenfassung

Der Aufsatz untersucht den optimalen Entwurf eines Regelsystems mit Vorhersageeigenschaften bei regellosen oder deterministischen Eingangsgrößen, wobei das Regelsignal entsprechenden Beschränkungen (Bedingungen) unterliegt. Die hier behandelten Beschränkungen des Regelsignals sind: a) der Betrag des Regelsignals ist konstant (Zweipunkt-Regelproblem) und b) der Betrag des Regelsignals ist kleiner oder gleich einer Konstanten.

Das grundlegende Optimierungsproblem wird hier recht allgemein als ein Problem der Entscheidungstheorie betrachtet. Auf Grund

der zur Verfügung stehenden Information ist zu entscheiden, welches der zulässigen Regelsignale der Strecke zuzuführen ist, um ein optimales Verhalten zu bekommen. Damit man das Problem auf ein solches der linearen Entscheidungstheorie zurückführen kann, wird das Optimum als Summe des quadratischen Fehlers definiert. Eine abschnittsweise lineare Grenze der Regelgröße wird gefunden, die sich zur Erzeugung des gewünschten Regelsignals verwenden läßt. Die Verwirklichung dieser Grenze des Regelsignals erweist sich als einfach und billig. Diese Methode läßt sich auf lineare, zeitunabhängige Strecken beliebiger Ordnung anwenden. Die Arbeit enthält einige Versuchsergebnisse, die diese Methoden bestätigen; so z. B. bei vorhandenem Rauschsignal die Regelung der Sprungantwort eines Systems mit einer durch ein Sättigungsglied begrenzten instabilen Regelstrecke zweiter Ordnung.

## Statement of the Problem

The problem being considered is the design of a predictor control system to achieve performance which is optimum in some sense<sup>1</sup>. Figure 1 is a block diagram of the system under consideration.

In the design problem, the performance index is first chosen. The control computer which will optimize the system with respect to that performance index must then be found. In this problem, the plant is assumed to be fixed and known, and the control signal is constrained in some way.

The control signal constraint of most interest is saturation. If  $m(t)$  is the control signal, this constraint is expressed as

$$|m(t)| \leq M \quad (1)$$

Systems with control signals governed by the constraint  $|m(t)| = M$  will be referred to as 'bang-bang' control systems.

The selection of a performance index is a compromise between what might be considered to be truly optimum performance and a practical solution. A practical solution is characterized by a straightforward design procedure and a simple control computer.

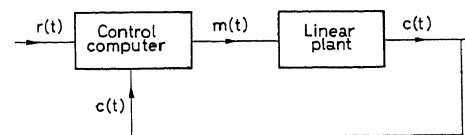


Figure 1

With this in mind, the error criterion chosen can best be explained in terms of a hypothetical system. This system has a piecewise-constant control signal (it can change values only at times  $t_k$ ) and is subjected to an input  $r(t)$ . It contains a control computer which employs a sub-optimal control law. This type of control law is sometimes referred to as sub-optimal since it does not take into account the fact that, in practice,

future decisions will be based on information not available at  $t = t_0$ . Instead, it generates only the initial value of the control signal arrived at through the resulting optimization and continually repeats the process.  $c^*(t)$  represents the output of the hypothetical system. The performance which is used to define the optimum is the minimization of

$$I = \sum_{k=1}^N (\bar{r}_k - c_k^*)^2 = \sum_{k=1}^N \bar{e}_k^2 \quad (2)$$

with respect to  $c_k^*$  [the notation used here is  $f(t_k) \equiv f_k$ ].

For deterministic inputs,  $\bar{r}_k = r_k$  and optimum performance corresponds to the minimization of the sum of squares of future errors of the hypothetical system. For stochastic inputs, minimization of

$$E \left[ \sum_{k=1}^N (r_k - c_k)^2 / P \right]$$

where  $P$  represents the history of  $r(t)$ , and  $c(t)$  is equivalent to minimization of  $I$  with respect to  $c_k^*$  where

$$\bar{r}_k \equiv E[r_k / P]$$

For Gaussian processes,  $\bar{r}_k$  can be found as the output of a least-squares predictor. The actual system under consideration differs from the hypothetical system previously considered in that the actual control signal is not constrained to be piecewise constant.

As a result of this difference between the actual and hypothetical systems, neither  $\bar{r}_k$  nor  $\bar{e}_k$  will in general appear at any time in the actual system. The error criterion is none the less reasonable and will form the basis of the work to follow.

### Formulation of Problem

$c(t)$  may be expressed as the sum of two signals. The first of these represents the output which would exist if the control signal were cut off at  $t = t_0$ . This is simply the solution of the homogenous differential equation describing the plant. This is referred to as  $c_0(t)$ . The second signal represents the change in output due to allowable changes in the control signal. This is referred to as  $\omega(t)$ . To express  $\omega(t)$ , it is advantageous to define a signal which represents the change in control signal at the various discrete times,  $t_k$ , when such a change is assumed permissible. This signal will be referred to as the switching signal,  $s_k$ , and will have the property

$$s_k = m_{k+1} - m_k \quad (3)$$

where  $m_k = m(t_k^-)$ . Thus  $m_k$  represents the control signal immediately before  $t_k$  and  $m_{k+1}$  represents the control signal immediately after.  $s_k$  thus represents the change in control signal occurring at  $t_k$ .

The plant being considered is described by either its impulse response,  $g(t)$ , or its step response,  $h(t)$ . The initial value of the step response is assumed to be zero.

Using the notation introduced above, one has

$$c_k^* = c_{k0} + \omega_k \quad (4)$$

where

$$\omega_k = \sum_{j=0}^{k-1} h_{k-j} s_j \quad (5)$$

$$\bar{e}_k = \bar{r}_k - c_k^* = \bar{r}_k - c_{k0} - \omega_k = z_k - \omega_k \quad (6)$$

where

$$z_k = \bar{r}_k - c_{k0} \quad (7)$$

Thus,  $z_k$  is determined by the behaviour of the reference input  $[r(t)]$ , and by the state of the system [this determines  $c_{k0}$ ]. Neither of these quantities can in any way be altered by the control computer.  $\omega_k$ , on the other hand, represents the change in output due to allowable changes in the control signal. The allowable values of  $\omega_k$  are completely determined by the plant and the control signal constraints. Thus, if one writes

$$I = \sum_{k=1}^N (\bar{r}_k - c_k^*)^2 = \sum_{k=1}^N (\bar{e}_k)^2 = \sum_{k=1}^N (z_k - \omega_k)^2$$

one notes that minimization of  $I$  with respect to  $c_k^*$  is entirely equivalent to minimization of  $I$  with respect to  $\omega_k$ .

The application of linear decision theory to the solution of the optimization problem at hand will now be demonstrated. Before proceeding further, it is necessary to introduce the following notation.

$\{m\} = \{m_1, m_2, \dots, m_N\}$  represents the control sequence. This notation is used to represent a signal,  $m(t)$ , which has the value  $m_1$  for  $0 < t < t_1$ ,  $m_2$  for  $t_1 < t < t_2$ , ... etc. The various elements of this control sequence are appropriately constrained.

$\{s\} = \{s_0, s_1, \dots, s_{N-1}\}$  represents the switching sequence. The constraints on the elements of this sequence are related to the control signal constraints through eqn (3).

$\{\omega\} = \{\omega_1, \omega_2, \dots, \omega_N\}$ . The constraints on the elements of the  $\omega$  sequence are related to the switching signal constraints through eqn (5). Corresponding to each control sequence there is an  $\omega$  sequence.

$\{z\} = \{z_1, z_2, \dots, z_N\}$ . The elements of this sequence are determined by the reference input and by the state of the plant.

**Theorem 1**—The sum of squared error criterion being employed in eqn (2) is a minimum-distance criterion<sup>2</sup>.

*Proof*—It is desired to minimize

$$I = \sum_{k=1}^N (z_k - \omega_k)^2$$

with respect to  $\omega_k$  subject to the appropriate control signal constraints. If one considers an  $N$ -dimensional space, a point,  $\omega$ , the elements of which are  $\omega_j$ , may be associated with each sequence  $\{\omega\}$ . The finite number of such points associated with a bang-bang control signal constraint will be referred to as pattern points. Similarly, corresponding to each sequence of  $z$ 's,  $\{z\}$ , there exists a point,  $z$ , in  $N$ -dimensional space.

Geometrically, it may be said that the minimization of  $I$  with respect to  $\omega$  corresponds to choosing the  $\omega$  point nearest to  $z$ . Corresponding to this  $\omega$  sequence, the optimum control sequence can then be found from eqn (5) and eqn (3).

**Theorem 2**—Linear decision functions are optimum for problems employing minimum-distance decision functions<sup>2</sup>.

*Proof*—A linear decision function has the property that it divides measurement space into regions such that each pair of regions is separated by one and only one hyperplane. Measurement space is the  $N$ -dimensional space in which the pattern points and the  $z$  points are plotted.

Figure 2(a) illustrates a minimum-distance criterion. Three pattern points are shown. A measurement  $Q$  is identified with that pattern point which is closest to it in an Euclidean sense.

Consider Pattern Points 1 and 2 (PP1 and PP2) and the hyperplane (line in this two-dimensional case)  $B_{12}$  which is the perpendicular bisector of the line joining PP1 and PP2 [Figure 2(b)]. Then the statement that a point  $Q$  is closer to PP1 than to PP2 is equivalent to the statement that the point lies on the 1 side of  $B_{12}$ . By constructing a hyperplane for every pair of pattern points, a linear decision function equivalent to the minimum-distance decision function is obtained.

The region associated with each pattern point will be referred to as a pattern class. Thus, associated with PP1 one has Pattern Class 1.

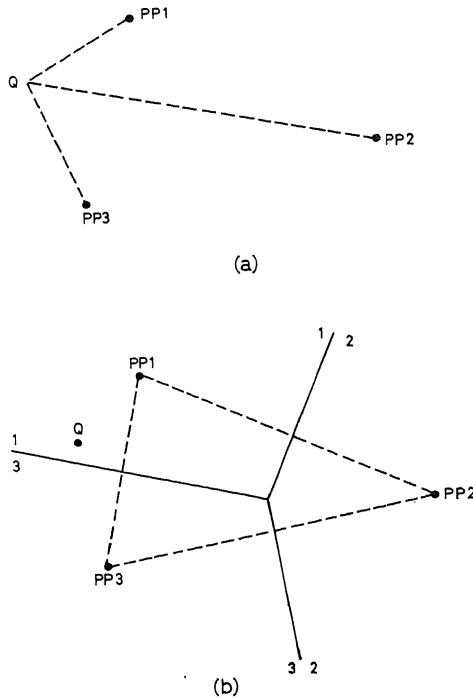


Figure 2. (a) Minimum-distance decision function; (b) linear decision function equivalent

### Representation of a Hyperplane in Normalized Form

Consider a hyperplane given by

$$\sum_{i=1}^N \alpha_i x_i + \alpha_0 = 0 \quad (8)$$

where

$$\sum_{i=1}^N \alpha_i^2 = 1 \quad (9)$$

Then the  $\alpha_i$ ,  $1 \leq i \leq N$ , are the direction cosines of the hyperplane, and  $\alpha_0$  is its distance from the origin. The distance,  $d$ , of a point  $P$  from this hyperplane is simply given by substituting the coordinates of the point into the normalized equation of the hyperplane;

$$d = \sum_{i=1}^N \alpha_i P_i + \alpha_0 \quad (10)$$

The point is on one side of the hyperplane if  $d$  is positive, and on the other side if  $d$  is negative. Which side of the hyperplane

is to be positive or negative is completely arbitrary, since multiplication of eqn (8) by  $-1$  changes the sign of  $d$  but does not change the hyperplane.

### Implementation of a Hyperplane

Figure 3 indicates the simplicity of implementation inherent in linear decision functions. The distance,  $d$ , of a point,  $z$ , from the hyperplane in question is proportional to the current  $i$ . The direction of current flow indicates the side of the hyperplane on which the point is located. The scale factor enables the use of realistic component values and signal levels.

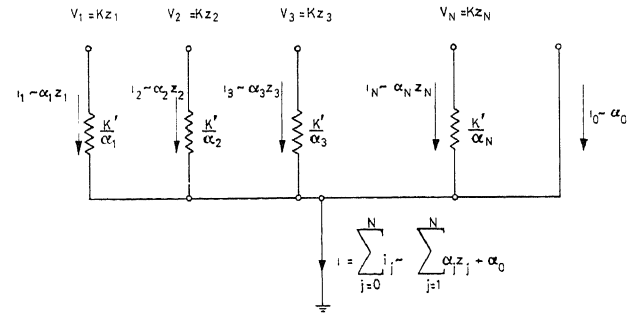


Figure 3. Implementation of a hyperplane

### The Classification Procedure

In order to be associated with Pattern Class  $i$ , a point must be on the  $i$  side of all boundaries  $B_{ij}$ ,  $j \neq i$ . If these boundaries are implemented, taking proper care of sign conventions, the distances from each of them can be combined in an AND fashion to arrive at the correct decision. In other words, if a positive sign is associated with points on the  $i$  side of  $B_{ij}$ , then all of the distances in question must be positive if the point is to be associated with Pattern Class  $i$ .

### Properties of $\omega$

#### Bang-bang Case

Since there are only two admissible values of  $m(t)$ , there are  $2^N$  possible control signals. Corresponding to each of these is an  $\omega$  sequence, the elements of which define the pattern points. For convenience  $M = \frac{1}{2}$  will be used. Any scale factor required can later be lumped with the plant.

As an example, consider the problem for  $N = 2$ . There are four possible control signals. The four possible control sequences, together with their corresponding switching sequences and pattern points, are listed in Table 1.

Table 1

	$a$	$b$	$c$	$d$
$\{m\}$	$\frac{1}{2}, \frac{1}{2}$	$\frac{1}{2}, -\frac{1}{2}$	$-\frac{1}{2}, -\frac{1}{2}$	$-\frac{1}{2}, \frac{1}{2}$
$\{s\}$	$\frac{1}{2}, 0$	$\frac{1}{2}, -1$	$-\frac{1}{2}, 0$	$-\frac{1}{2}, 1$
$\{\omega\}$	$h_1/2, h_2/2$	$h_1/2,$ $h_2/2 - h_1$	$-h_1/2,$ $-h_2/2$	$-h_1/2,$ $-h_2/2 + h_1$

These are shown in Figure 4 along with their associated pattern classes.

It is worth while, at this point, to consider in more detail the nature of the decision theory problem being investigated. The

optimum allowable control signal at  $t = 0^+$ ,  $m_1$ , is to be decided upon. In order to do this, one must consider the effect of the future control signals,  $m_2, m_3, \dots, m_N$  as well. One must thus select as the value of  $m_1$  that value which corresponds to the first element of the control sequence  $\{m_i\}$  which minimizes eqn (2). Since this decision process is carried out continuously, the optimum allowable control signal is continuously generated. The union of all pattern classes associated with the same value of  $m_1$  might be considered as a major pattern class, or contro

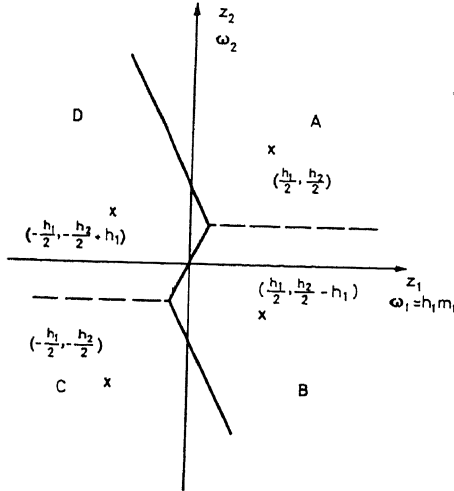


Figure 4. Pattern points and associated pattern classes for bang-bang case ( $N = 2$ )

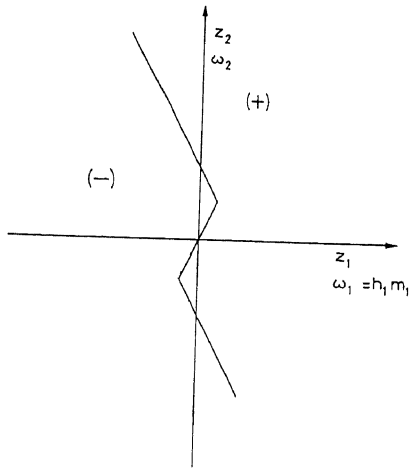


Figure 5. Control boundary for bang-bang case ( $N = 2$ )

class. As an example, the control classes for  $N = 2$  are shown in Figure 5. The union of Pattern Classes A and B defines the plus control class, since  $z$  points falling in either of these pattern classes have  $m_1 = +\frac{1}{2}$  associated with the optimum control signal. Similarly, Pattern Classes C and D define the minus control class, since  $z$  points falling in these regions correspond to  $m_1 = -\frac{1}{2}$ .

The piecewise linear boundary dividing these control classes will be referred to as the control boundary. Points lying on the plus side of the control boundary correspond to  $m_1 = +\frac{1}{2}$  while

points lying on the minus side correspond to  $m_1 = -\frac{1}{2}$ . Since the control boundary is determined by the pattern classes, the linear decision function defining the various pattern classes forms the heart of the control computer.

#### Saturation—Limited Case

For this case, admissible values of  $\omega$  lie in the parallelepiped defined by the pattern points associated with the bang-bang case\*. The interior of this parallelepiped will be referred to as the null region,  $R_0$ , since points falling in this region result in

$$I = \sum_{k=1}^N \bar{e}_k^2 = 0$$

For a given  $z$ , the problem then becomes that of finding the value of  $m_1$  corresponding to the point in the null region which is closest to  $z$  (since minimizing  $I$  with respect to  $\omega$  is equivalent to choosing that  $\omega$  which is closest to  $z$ ). This can best be explained by referring to the case for  $N = 2$  (see Figure 6). The optimum control signal,  $m_{1a}$ , corresponding to the point  $z_a$  is obtained from the  $\omega_1$  coordinate of the point in  $R_0$  which is nearest to  $z_a$ . The problem has thus been reduced to a geometry

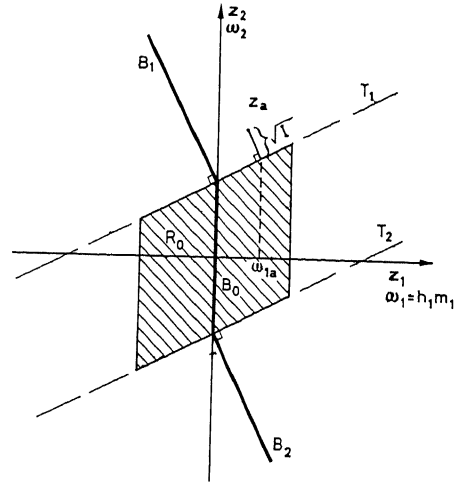


Figure 6. Null region, control and decision boundaries for  $N = 2$

problem. To facilitate the solution of this problem consider the Control Boundary  $B$  in Figure 6. It consists of three segments,  $B_0$ ,  $B_1$ , and  $B_2$ .  $m_1$  is proportional to the distance from the appropriate line segment. The proportionality factor is not the same in each case. Decision Boundaries  $T_1$  and  $T_2$  can be used to select the appropriate line segment (i.e., use  $B_1$  if  $z$  is above  $T_1$ ,  $B_2$  if  $z$  is below  $T_2$ , and  $B_0$  otherwise).

This represents an extension of what would normally be considered to be linear decision theory. Interest lies not only in which side of a control boundary  $z$  is situated, but how far it is from the boundary as well.

#### Implementation of $z_k$

Since  $z_k = \bar{r}_k - c_{k0}$ , its implementation can be considered in two parts.

\* The proof of this follows from a consideration of (1) and (5). It is included in the appendix of the dissertation upon which this discussion is based.

### Implementation of $\bar{r}_k$

In practice, since  $\bar{r}_k$  represents the predicted value of  $r(t)$ ,  $t_k$  seconds from the present, a predictor must be constructed to yield this. This paper does not deal with this subject.

### Implementation of $c_{k0}$

Since  $c_{k0}$  represents the solution of the homogeneous differential equation (HDE) describing the plant evaluated at time  $t_k$ , the implementation is straightforward. For an  $n$ th order plant,  $c_{k0}$  is a linear combination of the output and  $n-1$  derivatives (i.e., the state of the plant).

#### Example 1

$$G(s) = \frac{K}{s+a}$$

$$\text{HDE } \dot{c} + ac = 0$$

$$c_0(t) = b e^{-at} = c_0 e^{-at}$$

therefore

$$c_{k0} = c_0 e^{-at_k}$$

#### Example 2

$$G(s) = \frac{K}{(s+m_1)(s+m_2)}, \quad m_1 \neq m_2 \text{ and real}$$

$$\text{HDE } \ddot{c} + (m_1+m_2)\dot{c} + m_1m_2c = 0$$

therefore

$$c_0(t) = K_1 e^{-m_1 t} + K_2 e^{-m_2 t}$$

where

$$c_0 = K_1 + K_2$$

$$c'_0 = -m_1 K_1 - m_2 K_2$$

$$c_0(t) = \frac{m_2 c_0 + c'_0}{m_2 - m_1} e^{-m_1 t} + \frac{m_1 c_0 + c'_0}{m_1 - m_2} e^{-m_2 t}$$

$$\begin{aligned} c_{k0} &= \frac{m_2 c_0 + c'_0}{m_2 - m_1} e^{-m_1 t_k} + \frac{m_1 c_0 + c'_0}{m_1 - m_2} e^{-m_2 t_k} \\ &= \left[ \frac{m_2 e^{-m_1 t_k}}{m_2 - m_1} + \frac{m_1 e^{-m_2 t_k}}{m_1 - m_2} \right] + c'_0 \left[ \frac{e^{-m_1 t_k}}{m_2 - m_1} + \frac{e^{-m_2 t_k}}{m_1 - m_2} \right] \\ &= K_3 c_0 + K_4 c'_0 \end{aligned}$$

### Example of Bang-bang Predictor Control System Design

To clarify the application of the concepts and theorems discussed thus far, a complete design will be carried out for a first order plant with transfer function

$$G(s) = \frac{K_0 a}{s+a}$$

The step response,  $h(t)$ , can be evaluated and is given by

$$h(t) = K_0 (1 - e^{-at})$$

therefore

$$h_k = h(t_k) = K_0 (1 - e^{-at_k})$$

To be more specific, the case is considered for

$$N=2, \quad t_1 = T=0.692, \quad t_2 = 2T, \quad a=1$$

$r(t)$  will be a Gaussian random process obtained by passing white gaussian noise of zero mean and unit power spectral density through a simple low-pass RC filter of time constant  $1/b$ .

The four possible control signals, with their associated switching sequences and pattern points are listed in Table 1, where  $h_1 = 0.5 K_0$  and  $h_2 = 0.75 K_0$ .

The pattern points are shown plotted to scale in Figure 7. The control boundary is also shown. The plus control class can be implemented as the union of the regions to the right of boundary  $B_1$  or  $B_0 \cap B_2$ .

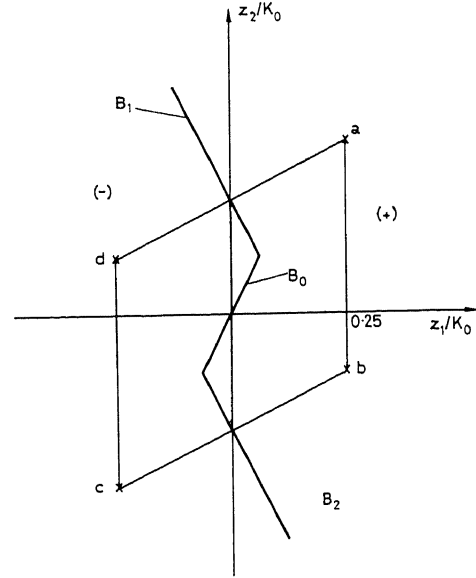


Figure 7. Pattern points and control boundary for bang-bang example

### Computation of Distance from Segment of Control Boundary

$d_i$  represents the distance of a point  $z$  from boundary  $B_i$ . The sign convention employed results in  $d_i > 0$  for points on the plus side of  $B_i$ .  $\alpha_{ij}$  represents the direction cosine in  $j$ th direction of boundary  $B_i$ .

#### (a) Computation for Boundary $B_1$

$$D_1 = |\overline{bc}| = [h_1^2 + (h_2 - h_1)^2]^{\frac{1}{2}} = 0.56 K_0$$

$$\alpha_{11} = \frac{h_1}{D_1} = 0.9 \quad \frac{1}{\alpha_{11}} = 1.1$$

$$\alpha_{12} = \frac{h_2 - h_1}{D_1} = 0.45 \quad \frac{1}{\alpha_{12}} = 2.2$$

$$\alpha_{10} = -\alpha_{12} \frac{h_1}{2} = -0.11 K_0$$

$$d_1 = \alpha_{11} z_1 + \alpha_{12} z_2 + \alpha_{10} = 0.9 z_1 + 0.45 z_2 - 0.11 K_0$$

#### (b) Computation for Boundary $B_0$

$$D_0 = |\overline{bd}| = [h_1^2 + (h_2 - 2h_1)^2]^{\frac{1}{2}} = 0.56 K_0$$

(it is a coincidence that  $D_0 = D_1$ )

$$\alpha_{01} = \frac{h_1}{D_0} = 0.9 \quad \frac{1}{\alpha_{01}} = 1.1$$

$$\alpha_{02} = \frac{h_2 - 2h_1}{D_0} = -0.45 \quad \frac{1}{\alpha_{02}} = -2.2$$

$$\alpha_{00} = 0$$

$$d_0 = \alpha_{01}z_1 + \alpha_{02}z_2 + \alpha_{00} = 0.9z_1 - 0.45z_2$$

(c) Computation for Boundary  $B_2$

$$D_2 = D_1 = 0.56 K_0$$

$$\alpha_{21} = \alpha_{11} = 0.9 \quad \frac{1}{\alpha_{21}} = 1.1$$

$$\alpha_{22} = \alpha_{12} = 0.45 \quad \frac{1}{\alpha_{12}} = 2.2$$

$$\alpha_{20} = -\alpha_{10} = 0.11 K_0$$

$$d_2 = \alpha_{21}z_1 + \alpha_{22}z_2 + \alpha_{20} = 0.9z_1 + 0.45z_2 + 0.11K_0$$

The required logic is shown in Figure 8.

Implementation of  $z_k$

1. The signal,  $r(t)$ , is a sample function from a random process with power spectral density

$$\Phi_r(\omega) = \frac{b^2}{\omega^2 + b^2}$$

Since  $\Phi_r(\omega)$  can be realized by exciting an RC filter with time constant  $1/b$  by white noise from a zero impedance source, the best predicting filter can be shown to be a simple attenuator.

Physically, this answer can be obtained by noting that, since the primary excitation voltage is white noise with zero mean, the best that can be said about the future is that the voltage across the capacitor  $n$  sec from now will simply be the result of the discharge of the present voltage. The transfer function of this predictor is simply

$$P_k = e^{-bt_k}, \text{ therefore } \bar{r}_k = r_0 e^{-bt_k}$$

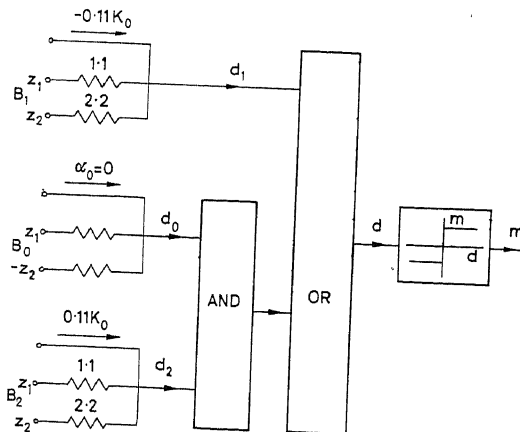


Figure 8. Control logic for bang-bang example

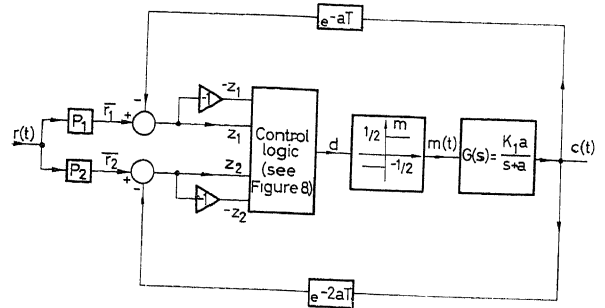


Figure 9. Complete block diagram for bang-bang example

Since  $r(t)$  is gaussian, this prediction represents the conditional expectation of  $r_k$  given the history of  $r(t)$ .

2. The signal,  $c_{k0}$ , is simply obtained from the solution of the homogeneous differential equation which describes the plant. For this case one has

$$c_{k0} = c e^{-at_k}$$

Thus

$$c_{10} = 0.5c$$

$$c_{20} = 0.25c$$

The overall block diagram for the system in question is shown in Figure 9.

#### Example of Predictor Control System Design Subject to Saturation

For purposes of comparison, the plant, input, and parameter values used to illustrate the design of predictor control systems subject to saturation will be the same as that which was used to illustrate the design of bang-bang predictor control systems.

$$G(s) = \frac{K_0 a}{s + a} \quad N=2, a=1, t_1=T=0.692, t_2=2T$$

The pattern points in this case are the same as in the bang-bang case (this will always be true). Figure 10 indicates  $R_0$  and

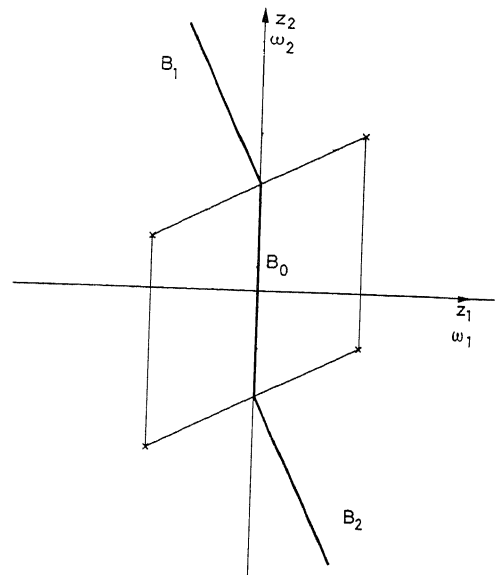


Figure 10. Null region and control boundary for saturation-limited example

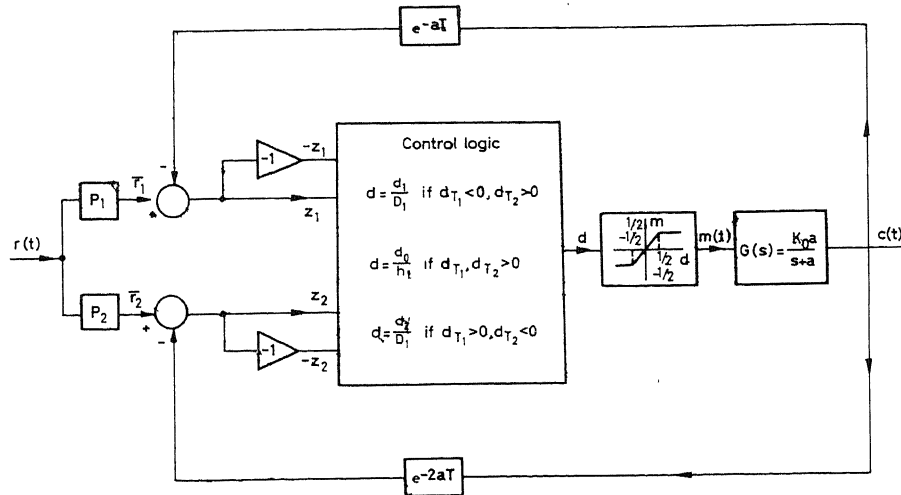


Figure 11. Block diagram for example of saturation-limited design

the new control boundary. The  $\alpha_i$  for the various control boundary segments have already been computed so the results will merely be presented here.

$$d_0 = z_1$$

$$d_1 = 0.9 z_1 + 0.45 z_2 - 0.11 K_0$$

$$d_2 = 0.9 z_1 + 0.45 z_2 + 0.11 K_0$$

In these expressions, the sign convention is such as to make  $d_i > 0$  for points on the plus side of the control boundary.

Decision Boundaries  $T_1$  and  $T_2$  are specified by the pattern points through which they pass. The resulting expressions for  $d_{T1}$  and  $d_{T2}$  are

$$d_{T1} = 0.45 z_1 - 0.9 z_2 + 0.225 K_0$$

$$d_{T2} = -0.45 z_1 + 0.9 z_2 + 0.225 K_0$$

In these expressions, the sign convention is such as to make  $d_{Ti} > 0$  for points on the  $R_0$  side of  $T_i$ .

The implementation of  $z_k$  is exactly the same as for the bang-bang example.

Figure 11 is a block diagram of the complete system.

### Results of Computer Simulation

An IBM 650 digital computer was employed to evaluate the step response of a predictor control system with a saturation-limited second order plant. Two specific cases were considered:

$$G_p(s) = \frac{1000}{s^2 + 2s + 100}; |m(t)| \leq \frac{1}{2}; \xi = 0.1, \omega_n = 10 \quad (11)$$

$$G_p(s) = \frac{1000}{s^2 - 2s + 100}; |m(t)| \leq \frac{1}{2}; \xi = -0.1, \omega_n = 10 \quad (12)$$

Case 1 represents a lightly damped but stable second-order plant. The unit step response (for  $N = 2$ ) is shown in Figure 12.

Case 2 represents an unstable second order plant. The unit step response (for  $N = 2$ ) is shown in Figure 13. This run was

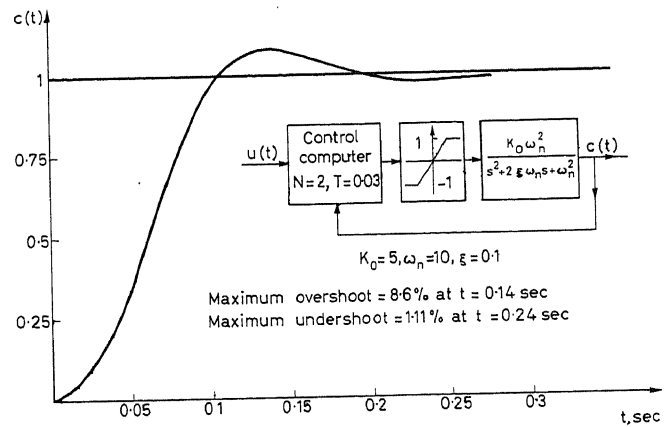


Figure 12. Step response of predictor control system with lightly-damped saturation-limited second order plant

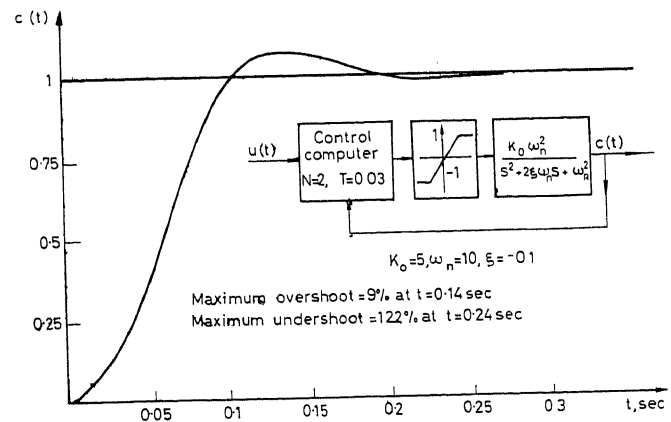


Figure 13. Step response of predictor control system with unstable saturation-limited second order plant

repeated with noise added to the plant input and the plant output remained essentially unchanged.

The work in this paper was performed at the Polytechnic Institute of Brooklyn as part of a Doctoral Dissertation and was supported in part by the Air Force Office of Scientific Research under grant number AFOSR-62-280.

## References

- <sup>1</sup> HORING, S. *On the Optimum Design of Predictor Control Systems*. Ph. D (EE) Dissertation, Polytech. Inst. Brooklyn (1962)
- <sup>2</sup> HIGHLEYMAN, W. H. *Linear Decision Functions with Application to Patterns Recognition*. D.E.E. Dissertation, Polytech. Inst. Brooklyn (1961)

## DISCUSSION

P. DORATO, *Polytechnic Institute of Brooklyn, Brooklyn, U.S.A.*

In order to understand some of the limitations of the method of design presented by the author, it is useful to outline the difference between the problem which must actually be solved and the problem which the author refers to as 'the hypothetical system'. The actual problem contains the following elements:

- (1) A linear plant with a continuous-time input  $m(t)$  and a continuous-time output  $c(t)$  with dynamics

$$c(t) = c_0(t) + \int_0^t h(t-\tau) m(\tau) d\tau$$

- (2) A performance index

$$I = E \left\{ \int_0^\infty e^2(t) dt | P \right\}$$

Where  $e(t) = c(t) - r(t)$ ,  $r(t)$  is, in general, some random process, and  $P$  represents the past data on which the conditional expectation  $E$  is conditioned.

- (3) Control input constraints

$$|m(t)| \leq \frac{1}{2}$$

The hypothetical problem contains the following elements:

- (a) A linear plant as above, with the only difference that now  $m(t)$  is constrained to be piece-wise constant, with changes in  $m(t)$  occurring only in intervals of  $T$  sec.

- (b) A performance index

$$I = E \{ [z(T) - \omega(T)]^2 + [z(2T) - \omega(2T)]^2 | P \}$$

where  $z(t) = c_0(t) - r(t)$  and where

$$\omega(t) = \int_0^t h(t-\tau) m(\tau) d\tau$$

- (c) Control input constraints

$$|m(t)| = \frac{1}{2} \text{ and } |m(t)| \leq 1$$

The author then makes use of some basic results in linear decision theory to solve the hypothetical problem. However, only the optimal control input at time  $t = 0$  is obtained, i.e.

$$m(0) = \frac{1}{2} \operatorname{sgn} \psi(0), \text{ for } |m(t)| = \frac{1}{2}$$

$$m(0) = \frac{1}{2} \operatorname{sat} \varphi(0), \text{ for } |m(t)| \leq \frac{1}{2}$$

The author's realization of the optimal controller for  $t = 0$  is quite simple (for the hypothetical system the control law is open-looped since the values of  $m(0)$  and  $m(t)$  depend only on data obtained for  $t \leq 0$ ).

The author then uses this control law obtained at time  $t = 0$  for the control of the actual problem for all time. The simplicity of the solution and the degree of similarity between the actual and hypothetical system make this design technique quite attractive. However the following problems do arise in the application of this technique to any given problem.

- (1a) The final control system may not even be stable, much less optimal.

- (2a) If the final system is stable, what is an optimum choice for the design variable  $T$ ?

Because of the complicated non-linear nature of the final control system it is difficult to study the above problems analytically. Thus in the application of the above design procedure to any particular system, some experimental studies should be made on the problem of stability and on the 'optimal' choice of  $T$ .

S. HORING, *in reply*

In reply to the comments made by Dr. Dorato, it should be noted that a detailed investigation of the stability of the resulting system has not yet been made. Neither has an optimum choice for the prediction time,  $T$ , been found. For the saturation-limited case, however, approximate procedures for overcoming these limitations do exist. Both procedures are based on quasi-linearization techniques which restrict consideration to points falling in the null region. Under this condition, the system behaves in a linear manner and linear stability techniques can be successfully applied. This is a 'small signal' approach, but appears to be reasonable. The selection of the prediction time,  $T$ , can also be made under the same conditions. If this is done, the procedure would merely require the specification of a performance index for the overall system and the selection of  $T$  so as to optimize the system with respect to this index.



# A Method of Optimal Control Prediction\*

F. B. GUL'KO and B. YA. KOGAN

## Summary

A method of optimal control for a class of any-order non-linear systems is described. The method is based on prediction of the optimal motion of all the system coordinates. Prediction is realized by formulating the optimal motion of a steadily decreasing number of elements of the plant at steadily increasing speed, using high-speed iterative computers. Some principles of the design of the predictors are presented, and experimental results for a fourth-order plant are given.

## Sommaire

Une méthode d'optimisation d'une classe de systèmes non-linéaires d'un ordre quelconque, est décrite. Cette méthode est basée sur la prédiction d'un déplacement optimal de toutes les coordonnées du système. Cette prédiction est réalisée en formulant la condition d'optimalité d'un nombre décroissant d'éléments du système à une vitesse croissante et en utilisant pour cela des systèmes numériques itératifs à très grande vitesse. Les principes de réalisation de tels systèmes de prédiction sont décrits et des résultats expérimentaux de leur application à un système du 4<sup>e</sup> ordre sont donnés.

## Zusammenfassung

Eine Methode für die zeitoptimale Regelung einer Klasse von nicht-linearen Systemen beliebiger Ordnung wird beschrieben. Die Methode beruht auf der Vorhersage der günstigsten Bewegung aller Systemkoordinaten. Die Vorhersage erfolgt durch Formulierung der günstigsten Bewegung für eine ständig abnehmende Anzahl von Gliedern der Strecke bei ständig zunehmender Geschwindigkeit; hierzu werden schnelle iterative Rechner verwendet. Die Arbeit enthält einige Ausführungsprinzipien für die Vorhersageeinrichtungen sowie Versuchsergebnisse für eine Strecke vierter Ordnung.

## Introduction

In view of the increasing demands which are made on the quality of automatic control processes, more and more use is being made of optimal control systems, particularly of a wide class of time-optimal systems.

The development of time-optimal systems is at present badly hampered by the difficulty involved in designing the controlling part of the system, which, except for the simplest cases of second-order linear plants, involves the use of multivariable functional generators, or of complex boundary-problem computers<sup>1</sup>.

Because of this, there has recently been a search for new approaches to the design of optimal control systems. In this connection mention should be made of the work of Coales and Noton<sup>2</sup>, who proposed that search of the switching moment of the control action should be realized on the basis of high-speed examination of a family of phase trajectories (future behaviours of the plant), on the assumption that this switching will take place at some future moment. Chestnut, Sollecito and Troutman

further developed this principle<sup>3</sup>, substituting for search of the switching moment a step-by-step analysis of a family of the phase trajectories, also obtained at high speed, on the assumption that switching of the control action has taken place at the current moment of time; the actual switching is executed when the predicted trajectory passes through the origin of the coordinates. Characteristic features of the above works are: (a) Prediction by repetitive computers of the set of future optimal behaviours (trajectories) of the plant, with verification of each trajectory to see whether it corresponds to the assigned boundary conditions; (b) the use in the control system of a logic for second-order plants, provided for not more than one switching of the control action. The latter confines the possible applications of these methods to second-order plants, or to plants reducible to the second order, having no supplementary constraints on the coordinates.

It is, however, possible to remove these constraints, at least for plants consisting of a number of series connected first-order elements, linear or with monotonic non-linearities (or plants reducible to such a form), by using the peculiarities of the structure of optimal processes in such systems. In this way it is possible to realize an optimal control system for a plant of the  $n$ th order, having an optimal controller for a plant of the  $(n-1)$ th order and a predictor. Applying the same principle in succession to plants of the  $(n-1)$ th,  $(n-2)$ th, ... orders, up to and including the second order, it is possible to construct an optimal control system for an  $n$ th order plant, the controlling part of which will consist of a set of predictors.

## Construction of Optimal Control Systems by Successive Reduction of the Order and Prediction

Considered here are plants described by a system of differential equations of the form:

$$\begin{aligned}\dot{x}_1 &= f_1(x_1, u) \\ \dot{x}_2 &= f_2(x_2, x_1) \\ &\dots\dots\dots \\ \dot{x}_n &= f_n(x_n, x_{n-1})\end{aligned}\quad (1)$$

where  $u = u(t)$  is the control action, while

$$|u(t)| \leq 1$$

All the functions  $f_i$  are assumed to be continuous and continuously differentiable with respect to  $x_i$  and  $x_{i-1}$ , while  $f_1$  is continuously differentiable with respect to  $u$ , and the partial derivatives  $\partial f_i / \partial x_{i-1}$  and  $\partial f_1 / \partial u$  do not change sign throughout the domain of variation of the variables in question.

Moreover, on some  $x_k$ , there can be imposed constraints of the form:

$$|x_k| \leq \bar{x}_k$$

\* General principles of this method were established by the authors in cooperation with Professor A. Ya. Lerner.

specifying the permissible domain of states of the system in the phase space. The problem is to synthesize a control system which effects the time-optimal shift of plant (1) from any initial state to any assigned equilibrium state.

To solve this problem, use will be made of a property of the structure of optimal processes in plants of type (1), namely that the trajectory of the optimal process consists of successive sections, on each of which the control corresponding to it coincides with the optimal control for a type (1) system having an order lower by a unity than the initial one. For example, if by some means it is possible to ensure, for a  $(n-1)$ th order plant [without the last element in (1)], a control action which in minimal time imparts to the coordinate  $x_{n-1}$ , an extremal value (taking into account the imposed constraints), and which then, in minimal time, transfers the plant to the assigned state (with respect to the  $n-1$  coordinate), and if, moreover, the coordinate  $x_n$  reaches the assigned value at the final moment, then the control action and corresponding trajectory of the whole system (1) are time-optimal. The proof of this is given by Gul'ko<sup>4</sup> (for the case when constraints of the  $\bar{x}_k$  type are lacking), where it is shown that such a control action satisfies Pontryagin's Maximum Principle<sup>5</sup>.

Figure 1 is the block diagram of an optimal control system based on these principles. The scheme consists of three main parts: the plant itself with an optimal controller [for the  $(n-1)$ th order], the predictor (P), and the logical gate (L).

The optimal controller in the plant assures optimal motion of the  $(n-1)$ th order plant towards the value of the coordinate  $x_{n-1}$  assigned by the logical gate L. The predictor is a high speed repetitive computer, which simulates the plant together with controller optimal for the  $n-1$  coordinate, the setting of which agrees with the assigned value of the coordinate  $x_{n-1}$ .

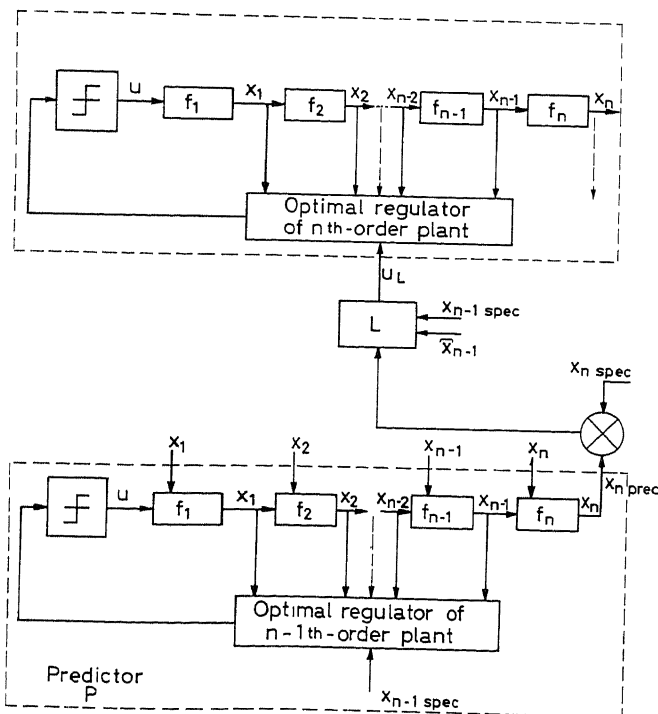


Figure 1. Block diagram of an automatic control system with prediction for one coordinate

Obtaining, at the beginning of every cycle of the solution, data on the state of the plant (the values of its current coordinates), the predictor computes the value which will be reached by the coordinate  $x_n$  if, starting from a given moment, the truncated  $(n-1)$  order system is brought to assigned equilibrium state in minimal time.

The output signal of the logical gate  $u_L$  is determined by the difference between the assigned value of the  $n$ th coordinate  $x_{n \text{ spec}}$  and its predicted value  $x_{n \text{ pred}}$ , from the equation:

$$u_L = \begin{cases} -\bar{x}_{n-1}^* \operatorname{sign}(x_{n \text{ spec}} - x_{n \text{ pred}}) \cdot \operatorname{sign} \frac{\partial f_n}{\partial x_{n-1}} & \text{when } x_{n \text{ pred}} \neq x_{n \text{ spec}} \\ x_{n-1 \text{ spec}} & \text{when } x_{n \text{ pred}} = x_{n \text{ spec}} \end{cases} \quad (2)$$

When  $x_{n \text{ spec}}$  is applied, a difference arises between  $x_{n \text{ spec}}$  and  $x_{n \text{ pred}}$ , as a result of which the logical gate L, in accordance with (2), sends to the optimal controller a signal for the variation of the coordinate  $x_{n-1}$ , (taking into account the sign of the difference). Moreover, the predictor continuously calculates the value which will be taken by the coordinate  $x_n$  if, at the given instant, the setting of the optimal controller is switched from  $\bar{x}_{n-1}$  to  $x_{n-1 \text{ spec}}$ . As soon as the value of  $x_{n \text{ pred}}$  reaches  $x_{n \text{ spec}}$ , the logical gate will bring about an actual change of the setting of the controller, after which the system will adopt the specified position, under the influence of the optimal controller. The predicted value  $x_{n \text{ pred}}$  remains unaltered over this interval of time.

The chain of reasoning employed for the synthesis of an optimal system of the  $n$ th order can be used to synthesize an  $(n-1)$ th order optimal controller for plant and predictor; that is, recourse is made to a system with an  $(n-2)$ th order optimal regulator and two predictors for the coordinates  $x_n$  and  $x_{n-1}$ . Naturally, in this case, the predictor which works out the future value of the coordinate  $x_{n-1}$ , and which itself forms part of the predictor for the coordinate  $x_n$ , must operate at a higher speed than the latter. Applying this method successively a further  $n-3$  times, one arrives at an optimal control system containing  $n-1$  predictors ( $P_1, P_2, \dots, P_{n-1}$ ) with their corresponding logical gates ( $L_1, L_2, \dots, L_{n-1}$ ), but containing no other optimal controllers (Figure 2). It is a characteristic feature of this system that the optimal nature of the calculated trajectories in any of the predictors is ensured by the presence in the make-up of any of them of other predictors which calculate the motion of a successively abbreviated number of elements at ever-increasing speed. Figure 3 is a block diagram illustrating the method of synthesis of the predictors.

To solve a tracking problem by the method described, recourse must be made to error equations, as was done by Coales and Noton<sup>2</sup>, and by Chestnut *et al.*<sup>3</sup>.

#### Optimal Control of a Fourth-order Plant

To illustrate the method, Figure 4 shows the example of a time-optimal control system for a fourth-order plant consisting of four integrating elements. The system contains three predictors:  $P_1, P_2$  and  $P_3$ . Let there be supplied to the system at some

\* If no constraint  $x_{n-1}$  is given, then the greatest value of the coordinate  $x_{n-1}$  that can be physically represented is introduced into the logic block in its place.

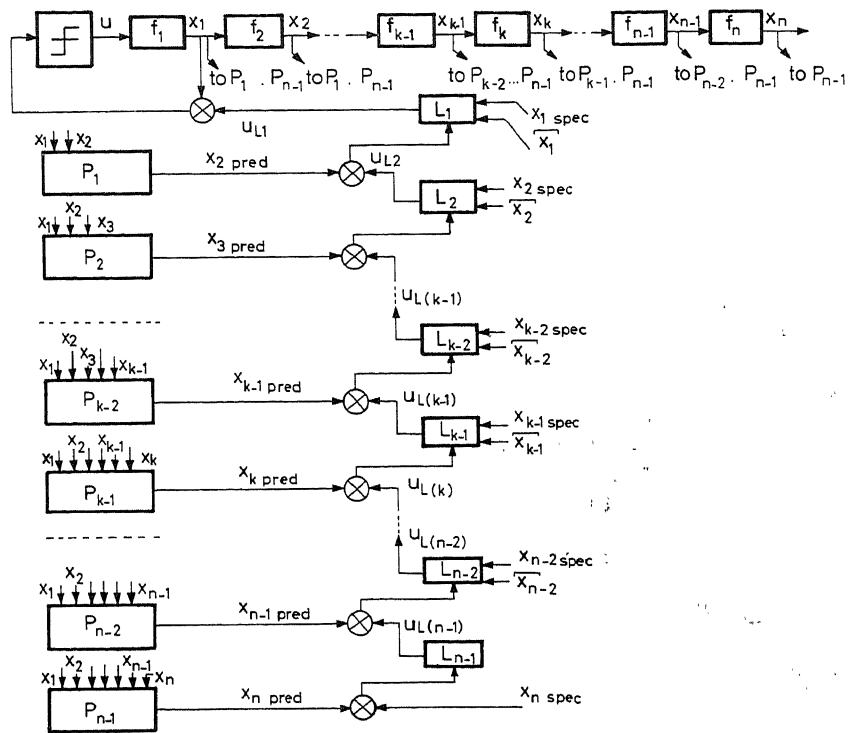


Figure 2. Block diagram of optimal control with prediction for  $n - 1$  coordinate (the general case)

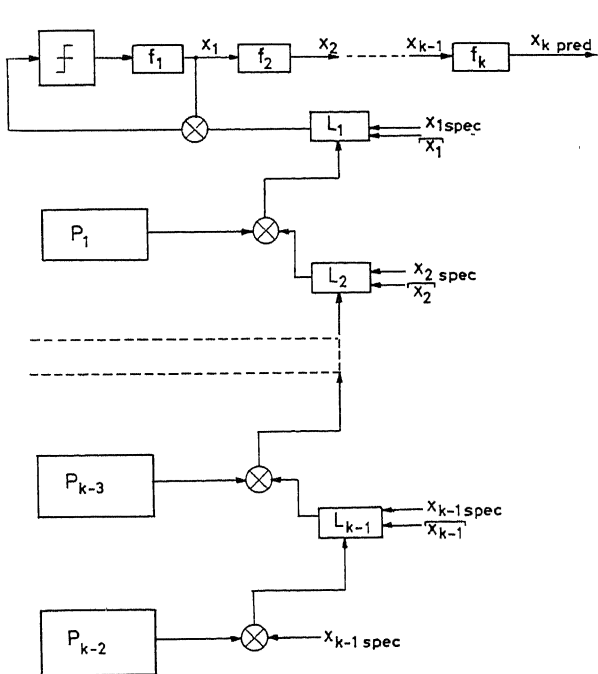


Figure 3. Block diagram of the  $(k - 1)$ th predictor

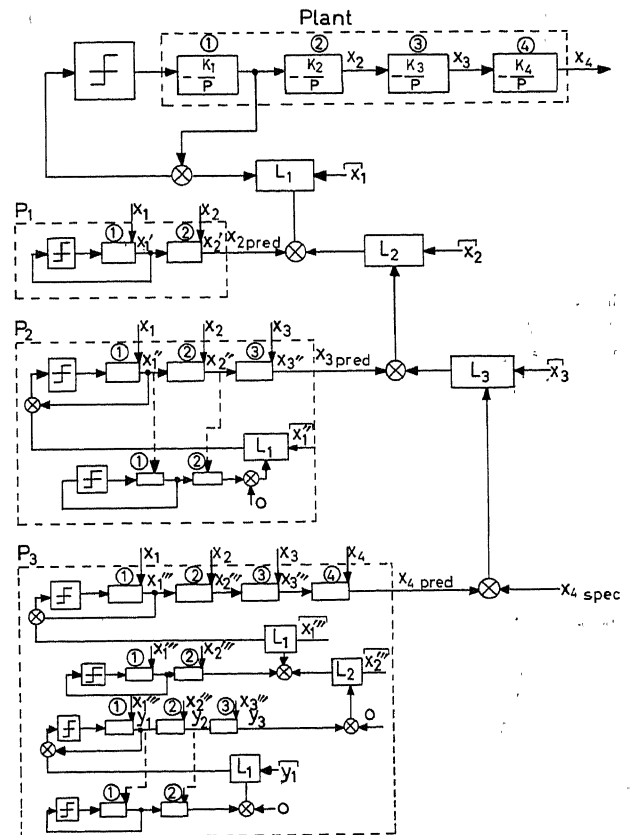


Figure 4. Automatic control system for fourth-degree plant

moment a 'specification' concerning the coordinate  $x_4$ . If at this moment the state of the system is such that, with optimal alteration of the coordinate  $x_3$  to the specified value corresponding to an equilibrium state (i.e., to zero), the coordinate  $x_4$  does not reach the specified magnitude, then as a result of the difference between  $x_{4 \text{ spec}}$  and  $x_{4 \text{ pred}}$  the logical gate  $L_3$  gives a signal  $u_{L3}$ , corresponding to the limit permissible value of the coordinate  $x_3$ , with the appropriate sign determined by the direction of the 'acceleration' of the coordinate  $x_4$ . If there is a difference between  $u_{L3}$  and  $x_{3 \text{ pred}}$ , logical gate  $L_2$  gives a 'specification'  $u_{L2}$  for the variation of the coordinate  $x_2$ , and logical gate  $L_1$ , under the same conditions, gives a specification  $u_{L1}$  for a variation of the coordinate  $x_1$  which switches the control relay. After this the system begins to 'accelerate' with maximum speed in the required direction, and at some moment  $t_1$  the value of  $x_{4 \text{ pred}}$  becomes equal to the given value  $x_{4 \text{ spec}}$ . Starting from this moment, the coordinate  $x_3$  must be zeroed with optimal rapidity, so logical gate  $L_3$  changes the command signal  $u_{L3}$  to zero. In the same way, logical gates  $L_2$  and  $L_1$  change the signs of  $u_{L2}$  and  $u_{L1}$ , and the first switching of the command relay takes place, after which the system starts to 'brake' with respect to the coordinate  $x_4$ . During the rest of the process,  $x_{4 \text{ pred}}$  will equal  $x_{4 \text{ spec}}$ . When the value of  $x_{3 \text{ pred}}$  reaches zero, gate  $L_2$  issues a command for optimal change of coordinate  $x_2$  to zero. This command enters the relay via gate  $L_1$ , and effects the second switching. After this,  $x_{3 \text{ pred}}$  remains equal to zero. The third switching occurs in the same way, when  $x_{2 \text{ pred}} = 0$ . This last interval ends when  $x_1$  reaches equilibrium value (zero). Thus by the end of the process  $x_3 = x_2 = x_1 = 0$  and  $x_4 = x_{4 \text{ spec}}$ . If, in the course of the process, the predicted value of some one of the coordinates, for example  $x_3$ , reaches a magnitude equal to the limit value prescribed for it by logical gate  $L_3$ , the specification for  $x_2$  will be switched to zero by the gate  $L_2$ , and the system will start 'braking' with respect to the coordinate  $x_3$ . At the end of this 'braking' process,  $x_3$  reaches its permissible value, and maintains it until the predicted value of  $x_4$  reaches the specified value. In this case the process will consist of a greater number of switchings, as follows from the theory of optimal control<sup>6</sup>. The results of simulating control processes by the proposed method are shown in the oscillograms of Figure 5 (a), (b), and (c), which show processes for the cases of no constraints on the phase coordinates, and the application of constraints on the third coordinate and the second and third coordinates together. Oscillograms of the outputs of the predictors  $P_3$  ( $x_{4 \text{ pred}}$ ) and  $P_2$  ( $x_{3 \text{ pred}}$ ) are also given.

If it is possible to realize optimal control of the first  $k$  elements of the plant ( $k = 2, 3, \dots$ ) in some other way, the number of predictors can be reduced by  $k - 1$ . The first of the remaining predictors must include a model of the corresponding optimal controller, the second a model of the first predictor, and so on, as has been described for the general case.

### Some Features of the Design of Repetitive Analogue Predictors

A predictor is a high-speed iterative computer, operating with acceleration of the processes (with respect to the plant). Because of the need for a high repetition rate, while the requirements for accuracy are relatively low, the use of analogue principles in the design of the predictors is most expedient.

The repetition rate is selected from considerations of the increment of the predicted quantity permissible, for reasons of

accuracy, in a cycle of the solution. The time scale is chosen with references to the repetition rate and the duration of the processes in the plant to be predicted by the device.

A predictor usually consists of the following main units: an analogue of the relay device supplying the control action; repetitive analogue computing elements (linear or non-linear), with a wide pass-band; a memory element, which stores information between cycles of the solution, and, finally, a control system, which provides the necessary sequence of switching operations.

The analogue relay element is usually a flip-flop or operational amplifier with a limiter in the feedback circuit or on the output.

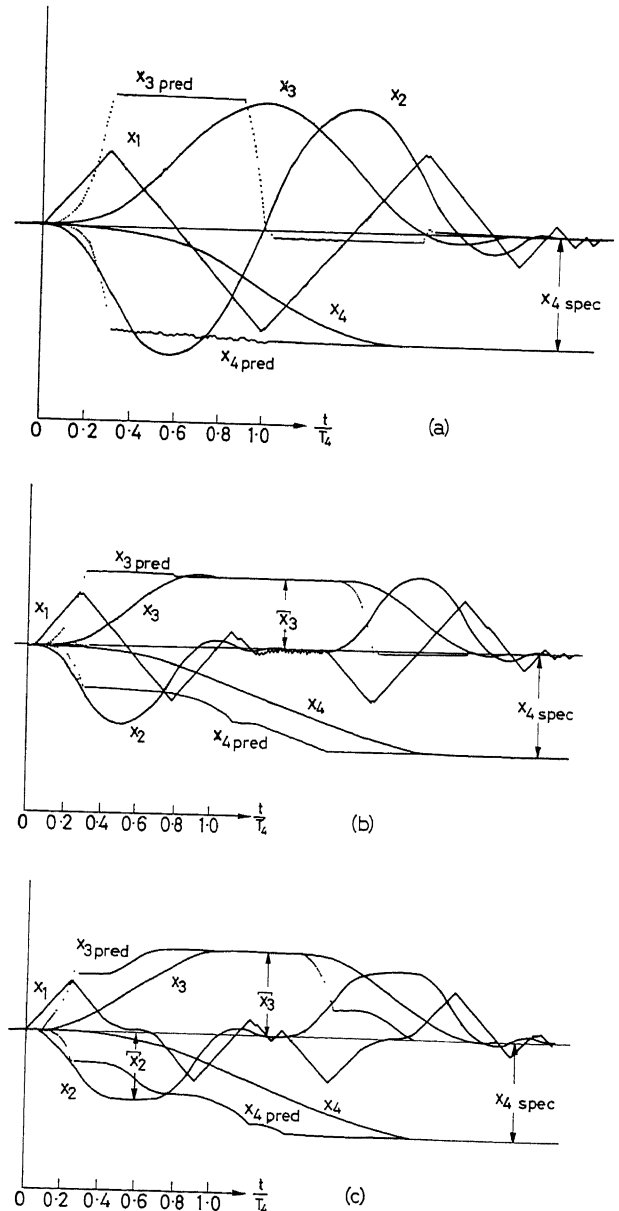


Figure 5. Oscillograms of transient processes in the optimal control system of a fourth-degree plant:

(a) with no constraint on the phase coordinates; (b) with a constraint on the coordinate  $x_3$ ; and (c) with a constraint on the coordinates  $x_2$  and  $x_3$ .

The amplifiers for the computing elements must ensure the desired accuracy of operation, and must have low drift. For this purpose, according to Polonnikov<sup>7</sup>, the most suitable is a direct current amplifier with a zero drift compensation network based on the ideas of Prinz. Figure 6 gives the structural scheme of a scale computing element and integrator. Here in the RESET cycle all the switches are closed, and because of the full negative feedback the capacitor  $C_k$  is charged up to the drift voltage at the amplifier output. In the 'solution' cycle the switches are opened, and the compensating voltage across the capacitor  $C_k$  is connected in series with the voltage at the summing point. Ordinary switches of the bridge type are used.

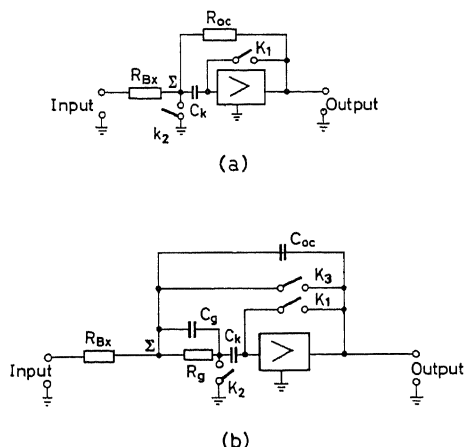


Figure 6. Zero drift compensation circuit of computing elements. (a) Scale computing element, and (b) integrator

The memory element stores the solution at the end of an operational cycle for the duration of the RESET period of the next solution cycle. It can be constructed from a combination of operational amplifier, memory capacitors and switches, for example.

The control system produces clock pulses which control the mode of operation of the switch, and can be constructed using conventional multivibrators.

## Conclusions

The theoretical possibility of constructing time-optimal control systems for  $n$ th order plants has been demonstrated, using a set of predictors as the optimal controllers.

Such optimal control systems are synthesized according to a proposed way, on the basis of a mathematical description of the plant. Elaborate calculations are not required, and the adjustment of the control system is simplified.

The predictors used in the control system can also function as 'advisors' to the operators in the case of manual control. For third-order and fourth-order plants, the method described can be realized with the aid of a comparatively simple apparatus, which can be built with the technical means now available.

The extension of the method to other, more complex, optimal control problems will require further investigation of the structural features of optimal processes, and the development of very high-speed and reliable means of mathematical simulation.

## References

- 1 FELDBAUM, A. A. *Computers in Automatic Systems*. Ch. 13. 1959. Moscow; Fizmatgiz
- 2 COALES, J. F., and NOTON, A. R. M. An on-off servomechanism with predicted changeover. *Proc. Instn elect. Engrs Pt B*, 103, No. 10 (1956), 449-462
- 3 CHESTNUT, H., SOLLECITO, W. E., and TROUTMAN, P. H. Predictive control system application. *Appl. and Ind.* 55 (1961), 128-134
- 4 GULKO, F. B. A special feature of the structure of optimal processes. *Izvest. Akad. Nauk USSR, OTN, Teshnicheskaya Kibernetika*, No. 1 (1963), 91-97.
- 5 PONTRYAGIN, L. S., BOLTYANSKII, V. G., GAMKRELIDZE, R. V., and MISHCHENKO, E. F. *A Mathematical Theory of Optimal Processes*. 1961. Moscow; Fizmatgiz
- 6 LERNER, A. YA. Maximal quick response of automatic control systems. *Automat. Telemekh.* 15, No. 6 (1954), 461-477
- 7 POLONNIKOV, D. F. *Operational Amplifiers for Repetitive Computers*. 1961. Moscow; 2nd All-Union Seminar-Conference on the Theory and Methods of Mathematical Simulation Moscow 1961

## DISCUSSION

I. F. COALES, *The University, Cambridge*

This is an interesting paper giving a rigorous method for designing a time-optimal control of an  $n$ th order system using predictive control, but, intuitively, with random inputs. I cannot see any need to go beyond prediction of the first derivative of error as was proposed in our original paper<sup>1</sup>.

Experiments since carried out in various places have shown that in a number of practical cases the improvement resulting from predicting the second derivative as well as the first has resulted in only an insignificant improvement in performance.

There is plenty of practical evidence that predictive control can be used successfully for higher-order systems, both linear and non-linear, provided that the fast model of the plant is a sufficiently close approximation to the actual plant. Of course, when one comes to higher-order systems, care has to be taken to ensure that the iterative process is convergent, since otherwise the system may become unstable.

The object of work on predictive control is, in my view, to find iterative methods of control for complex systems with randomly chang-

ing inputs and parameters which will give sub-optimal solutions of adequate performance that are relatively easily realized.

## Reference

- 1 COALES, J. F. and NOTON, A. R. M. An on-off servomechanism with predicted changeover. *Proc. Inst. elect. Engrs Pt. D*, 103, No. 10 (1956) 449-462

B. YA. KOGAN, *in reply*

Optimal-response control of high-order systems based on the logic of optimal control of second-order systems can be effected only for those particular cases when there are two dominant time constants in the system, and the remaining time constants are so small that they can be considered parasitic, i.e. for systems practically reducible to second-order systems.

In our paper we pose the problem of how to achieve optimal-response control with the aid of predictors for high-order systems not reducible to second-order systems.

In contrast to the method of Coales and Noton<sup>1</sup>, the method explained in the paper does not require an extra adaptive system, the stability of operation of which depends on the order of the equation of the plant, and therefore with a rise in the order of the system, difficulties do not arise with the stability of the loop producing the switching moment.

Here difficulties arise in connection with the increase in the volume of equipment and the requirements for its speed of response.

It has recently been possible to prove that the method explained in the paper can also be extended to plants containing oscillatory elements<sup>2</sup>.

I must agree that in further studies of control using predictors, attention must be paid to complex systems controlled by randomly

varying input signals against a background of interference, and also to the creation of close to optimal control systems with the aid of predictors.

This will make it possible to simplify control equipment considerably without noticeable impairment of the quality of operation.

#### References

- <sup>1</sup> COALES, J. F. and NOTON, A. R. M. *Inst. elect. Engrs* 103, Pt. B, No. 10 (1956)
- <sup>2</sup> MIKHAILOV, N. N. and NOVOSELTSEV, ZH. A. Optimal transient processes in a third-order system, containing an oscillatory element and an integrator, and their realisation using prediction. *Tekhn. Kibernetika*, No. 6 (1963)

# Optimalization of Non-linear Random Control Processes

R. KULIKOWSKI\*

## Summary

There are known controlled systems such as chemical plants or airplanes whose differential equations are not known completely to the controller because of environmental changes, ageing, etc. In many systems of this kind the best which can be accomplished is to construct a multistage optimizing process which converges to the optimum control. In the case of zero-memory, non-linear plant processes of this kind can be constructed if the gradient of the performance measure is known or can be determined experimentally.

In this paper an extension of this method is considered for the case of non-linear plants having memory and changing randomly in time. In the two introductory parts of the paper the assumptions and the necessary and sufficient conditions of optimality are formulated. It is shown that at any stage of optimalization the best change of the input signal should be adjoint to the mean value of the gradient of the performance measure. Then an optimizing process, based on the so-called fixed point theorem, is constructed. It is shown that for certain classes of non-linear random plants all the necessary information about the generalized gradient can be obtained experimentally. As an example an optimizing process which minimizes the cost of input energy and maximizes the output gain of a non-linear plant has been constructed and is discussed.

## Sommaire

On sait qu'il existe des systèmes à optimiser, tels des usines chimiques ou des avions, dont les équations différentielles caractérisant leur comportement ne peuvent être complètement prédéterminées car elles dépendent de circonstances extérieures au système et varient continuellement. La meilleure façon de procéder est alors d'utiliser un procédé d'optimisation par approximations successives qui convergent vers un régime optimal. Des dispositifs d'optimisation dépourvus de mémoire peuvent être alors réalisés si le gradient caractérisant la performance à atteindre est connu ou peu être déterminé expérimentalement. Ce rapport considère une extension de cette méthode pour le cas de systèmes non-linéaires avec mémoire, et évoluant continuellement de façon aléatoire.

Les 2 premières parties de ce rapport indiquent les hypothèses et les conditions nécessaires et suffisantes d'une telle optimisation. Il est montré qu'à chaque étape de cette optimisation, la meilleure variation du signal d'entrée doit être ajoutée à la valeur moyenne du gradient de la mesure de la performance. Un processus d'optimisation basé sur le théorème du 'point fixe', peut être alors réalisé. Le rapport montre que pour certaines installations non-linéaires de caractère aléatoire, les informations nécessaires pour la détermination du gradient généralisé, caractéristique des performances à atteindre, peut être obtenu expérimentalement. Il décrit et discute un exemple de processus d'optimisation destiné à minimaliser le coût de l'énergie d'entrée, et à maximaliser le gain de sortie d'un système non-linéaire.

## Zusammenfassung

Es gibt Regelstrecken, so z. B. chemische Anlagen oder Flugzeuge, deren Differentialgleichungen für die Regelung nicht vollständig bekannt sind, da Änderungen der Umgebungsbedingungen, Alterung

usw. vorliegen. Für viele solcher Systeme ist der Aufbau eines mehrstufigen Optimierungsprozesses, der nach einer optimalen Regelung strebt, das Beste, was sich erreichen läßt. Im Falle nicht-linearer Regelstrecken ohne Gedächtnis können Prozesse dieser Art aufgebaut werden, wenn der Gradient des Gütemaßes bekannt oder experimentell bestimmbar ist. In diesem Beitrag wird eine Erweiterung dieser Methode betrachtet, und zwar für den Fall, daß die nichtlinearen Regelstrecken auch Glieder mit Gedächtnis enthalten und sich im Laufe der Zeit zufallsbedingt ändern. In den zwei einführenden Teilen des Beitrages werden die Voraussetzungen sowie die notwendigen und hinreichenden Bedingungen für das Optimum formuliert. Es wird gezeigt, daß bei jeder Stufe der Optimierung die beste Änderung des Eingangssignales adjungiert zum Mittelwert des Gradienten des Gütemaßes sein sollte.

Ein Verfahren zur Optimierung wird auf der Grundlage des sogenannten „Festpunktheoremes“ entwickelt. Es zeigt sich, daß für bestimmte Klassen nicht-linearer zufallsgestörter Regelstrecken alle notwendigen Informationen über den verallgemeinerten Gradienten experimentell bestimmt werden können. Als Beispiel wird ein Optimierungsverfahren abgeleitet und diskutiert, der auf ein Minimum der Kosten der Eingangsenergie und auf ein Ertragsmaximum bei einer nicht-linearen Strecke führt.

## Introduction

In the theory of optimum control systems it is usually assumed that the plant differential or operator equations are completely known to the controller. In such cases, through the application of known optimalization techniques, the optimum control signal can be derived by an analogue or digital computer and applied to the plant during any time interval. However, there are known systems such as chemical plants and aircraft whose differential equations are not known completely to the controller because of environmental changes, ageing, etc. In many systems of this kind the best that can be accomplished is to construct a multistage optimizing process which converges to the optimum control. All the necessary information for the construction of such a process can be obtained by observing outputs of the plant at every stage for known inputs. Applying this approach to non-linear, zero-memory plants, the gradient of the performance measure can be determined and known iteration methods, based on the gradient concept (such as steepest descent, non-linear programming, contracting iterations, Newton method, etc.) can be applied<sup>1</sup>.

If it is desired to extend these methods for the case of non-linear and random plants having memory (i.e., possessing inertial elements), with the object of obtaining a stable optimizing process, one should first define and determine experimentally the generalized gradient of the performance measure and then construct the convergent iteration process. It will be shown that these problems can be solved successfully, at least in the case of certain classes of non-linear, inertial, random plants, by using some concepts of non-linear and probabilistic functional

\* This research was partially supported by the National Science Foundation under Grant NSF-G-14514.

analysis. However, since the author realizes that one of the main purposes of a short technical paper is to present the arguments and results in a form which is understandable for the majority of engineers, an attempt has been made to avoid abstract formulations. The more delicate, formal questions are therefore explained in *Remarks I, II, III* which can be omitted during the first reading.

### Assumptions

(1) It will be assumed that the controller generates signals  $x(t)$  which may be subject to certain constraints, such as volume or energy constraints, i. e.

$$\int_0^T |x(t)|^p dt \leq L = \text{const. where } p = 1, 2 \quad (1)$$

or amplitude constraints, i. e.

$$\max |x(t)| \leq M = \text{const. etc.} \quad (2)$$

The controller can observe the output  $y(t)$  of the plant for every  $x(t)$  applied to the input by the feedback loop (see Figure 1).

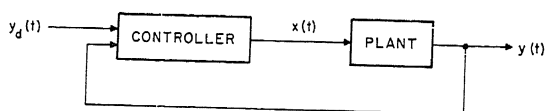


Figure 1. The optimizing control system

(2) We shall also assume that the form of the output-input relation of the controlled system can be described with sufficient accuracy by a non-linear, twice differentiable, integral operator. This operator for example, may be of the polynomial type:

$$y = A(x) = A_0(t) + \sum_{i=1}^{\infty} \int_0^T \dots \int_0^T k_i(t; \tau_1 \dots \tau_i) x(\tau_1) \dots x(\tau_i) d\tau_1 \dots d\tau_i \quad (3)$$

where the kernels  $k_i$  and the function  $A_0(t)$  are generally unknown to the controller.

The differential  $dA(x, h)$  of the operator  $A(x)$ , which is an extension of the usual concept of the differential of a function, can be defined as

$$dA(x, h) = \lim_{\gamma \rightarrow 0} \frac{1}{\gamma} \{A[x(t) + \gamma h(t)] - A[x(t)]\} = \frac{d}{d\gamma} A[x(t) + \gamma h(t)]_{\gamma=0} \quad (4)$$

where  $h(t)$  is an arbitrary function subject to the same constraints as  $x(t)$ . We assume also that it is possible to determine the approximate value of (4) experimentally by observing the outputs of the plant for  $x(t)$  and  $x(t) + \gamma h(t)$  and computing<sup>3</sup>

$$\frac{1}{\gamma} \{A[x(t) + \gamma h(t)] - A[x(t)]\} \approx dA(x, h) \quad (5)$$

where  $\gamma$  is a sufficiently small number.

(3) It is assumed that a performance measure  $F(x)$  is given

$$F(x) = \int_0^T G[x, y, y_d] dt \quad (6)$$

where  $G[x, y, y_d]$  is a known, twice differentiable function of the arguments  $x, y$ .

As an example, consider a chemical plant (for instance, a reactor, distillation column, etc.) described by the positive operator  $A(x)$  [which is non-negative for any  $x(t)$ ]. The amount of steam, fuel or electrical energy delivered to the plant within the time interval  $[0, T]$  will be equal to  $\int_0^T |x(t)|^p dt$ , where  $p = 1$  or 2. The output product obtained in time  $T$  from the plant will be  $\int_0^T A(x) dt$ . Then as the cost of running the plant in the time  $T$ , we can take the performance expression

$$F(x) = \lambda_1 \int_0^T |x(t)|^p dt - \lambda_2 \int_0^T A(x) dt \quad (7)$$

where  $\lambda_1, \lambda_2$  = positive coefficients which express the cost in the accepted currency.

As the next example consideration can be given to an autopilot-controller, which minimizes the integral error between the desired  $y_d(t)$  and the actual  $y = A(x)$  path angle of an aircraft:

$$F(x) = \lambda \int_0^T w(t) |y_d(t) - A(x)|^p dt \quad (8)$$

where  $w(t)$  = the given weighting function and  $x(t)$  is subject to the amplitude or integral constraints (2) or (1).

In the case where we want to minimize the final deflection  $\varepsilon(T) = y_d(T) - y(T)$  and its derivatives  $\varepsilon^{(i)}(t)|_{t=T}$ , i. e., when

$$F(x) = \lambda \sum_{i=0}^n \lambda_i \left| \frac{d^i}{dt^i} \varepsilon(t) \right|_{t=T} \quad (9)$$

the weighting function

$$w(t) = \sum_{i=0}^n \lambda_i (-1)^i \delta^{(i)}(T-t) \quad \text{and} \quad p = 1$$

should be substituted into (8). The problem becomes more complicated when one wants to minimize the time  $T$  subject to the constraints  $\varepsilon^{(i)}(T) = 0$ , and (1) or (2).

(4) In the general case  $A(x)$  may be a random operator, i. e., for the same  $x(t)$  it may be the case that  $y(t) = A(x)$  is a random function. Therefore, in the performance measures (7), (8), (9) the expected values will be assumed, i. e.  $E\{A(x)\}$  instead of  $A(x)$ .

In many cases  $y_d(t)$  is not known *a priori* and the transient term  $A_0(t)$  caused by the non-zero initial conditions of  $A(x)$  is not known as well. Therefore in the case of (8), (9) it will be assumed that the function  $y_d(t) - A_0(t)$  can be predicted so that it will be known, at least approximately, in the interval  $[0, T]$  and  $A(x)$  will not depend on the initial conditions. In the case of (7) the output  $y(t)$  is the sum of the processes due to  $x(t)$  acting within  $[0, T]$  and  $x(t)$  acting in the past, and this last term contributes to the output.

Now the goal can be formulated, and it is necessary to find a control signal  $x(t)$  which will minimize the performance measure  $F(x)$ . In order to solve this problem one has to determine the conditions of optimality and construct the optimizing process which will converge to the best  $x(t)$ .

*Remark I.* Speaking more precisely, one wants to minimize the twice-weakly differentiable functional  $F(x)$ , determined on the open or closed sphere of  $L^p[0, T]$  space

$$\|x\| = \left\{ \int_0^T |x(t)|^p dt \right\}^{1/p} \leq R, \quad 1 \leq p$$



The norm  $\|x\|$  in the space  $L^\infty$  should be defined as the so-called 'essential maximum' or

$$\|x\| = \inf_E \left\{ \sup_{t \in [0, T] - E} |x(t)| \right\}, \quad \text{mes } E = 0$$

which is, roughly, equivalent to (2). The functional (9) should be regarded as the so-called Schwartz distribution or generalized function. The  $\delta^{(i)}(t)$  functions can then be defined as the limits of weakly converging linear functionals.

The concept of a random operator is based on the notion of the so-called generalized random variable<sup>2</sup>. Usually, in control theory, random phenomena are described by random numbers or stochastic processes, which are, roughly, random numbers for any fixed time moments. It is known that the random numbers can be defined in the axiomatic way as the mapping of the space of events into the space of real numbers. It is possible to extend the notion of random numbers to the generalized random variable, which is a Borel measurable mapping of the space of events into some topological or metric space (in our case only, a sphere of  $L^p[0, T]$  space). More precisely<sup>2</sup>, let  $(\Omega, S)$  be a measurable space and  $X$  a non-empty metric space with the  $\sigma$ -algebra  $Z$  of all Borel subsets of the space  $X$ . Then the mapping  $V$  of the space  $\Omega$  into  $X$  is called a generalized random variable if the inverse image under the mapping  $V$  of each Borel set  $B$  belongs to the  $\sigma$ -algebra  $S$ , or in symbols, if  $\{\omega : V(\omega) \in B\} : B \in Z\} \subset S$ .

The random operator, which can be denoted by  $A(\omega, x)$ ,  $\omega \in \Omega$ , can be defined as the operator which for every fixed  $x$  is a generalized random variable.

The expected value of  $A(\omega, x)$  can be defined as the Bochner integral over the space  $\Omega$ :

$$E\{A(x)\} = \Delta \int_{\Omega} A(\omega, x) d\mu(\omega)$$

where  $\mu$  is the probability measure, i.e. a non-negative, countable, additive, real set function with the property  $\mu(\Omega) = 1$ . It is assumed that  $E\{A\}$  exists and the expectation sign will be treated as a linear operator acting from the random variable space into the output signal space  $Y$ .

### Conditions of Optimality

When  $x(t)$  is optimum, any variation  $\gamma h(t)$  of  $x(t)$  should not decrease  $F(x)$ . For example, taking  $G[x, y, y_d] = \lambda x^2(t) - g[y, y_d]$  one can express this condition in the form:

$$\begin{aligned} dF(x, h) &= \frac{d}{d\gamma} F[x + \gamma h] \Big|_{\gamma=0} \\ &= 2\lambda \int_0^T x(t) h(t) dt - \int_0^T \frac{dg}{dy}[y, y_d] \frac{d}{d\gamma} A[x + \gamma h] \Big|_{\gamma=0} dt \\ &= 2\lambda \int_0^T x(t) h(t) dt - \int_0^T g'[y, y_d] dA(x, h) dt = 0 \end{aligned} \quad (10)$$

assuming that the second differential  $d^2/d\gamma^2 F[x + \gamma h] \Big|_{\gamma=0}$  is positive for all  $h(t)$ .

It is more convenient to formulate this condition in a form which does not depend upon the arbitrary function  $h(t)$ . If an example is taken of the operator

$$A(x) = \int_0^T k_1(t-\tau) \left[ \int_0^T k_2(\tau-\tau_1) x(\tau_1) d\tau_1 \right]^n d\tau \quad (11)$$

which has the following differential

$$\boxed{\text{Eqn (12)}}^*$$

one can substitute (12) into (10) and by interchanging the integration order there is obtained

$$dF(x, h) = \int_0^T h(t) dt \{ 2\lambda x(t) - dA^*(x, g') \} = 0$$

where

$$\boxed{\text{Eqn (13)}}^\dagger$$

Then it can be observed that  $dF(x, h) = 0$  for every  $h(t)$  if

$$f(x) = 2\lambda x(t) - dA^*(x, g') = 0 \quad (14)$$

The operator  $f(x)$  will be called the gradient of  $F(x)$ ,  $[f(x) = \text{grad } F(x)]$  because it can be regarded as a generalization of the notion of gradient as commonly thought of in analytic geometry.

When the gradient  $f(x)$  of  $F(x)$  in the neighbourhood of a certain  $x(t) = x_0(t)$  is known, it is possible to express the decrease of  $F(x)$  along the trajectory  $x_0(t) + \alpha[x(t) - x_0(t)]$  (where  $\alpha$  is changing from  $\alpha = 0$  up to  $\alpha = 1$ ) by the mean value

$$\bar{f}(x_0, x) = \int_0^1 d\alpha f\{x_0(t) + \alpha[x(t) - x_0(t)]\}$$

of the gradient along this trajectory. Indeed, one obtains (see *Remark II*) the following inequality:

$$\begin{aligned} |F(x_0) - F(x)| &\leq \left\{ \int_0^T |\bar{f}(x_0, x)|^q dt \right\}^{1/q} \left\{ \int_0^T |x_0(t) - x(t)|^p dt \right\}^{1/p} \end{aligned} \quad (15)$$

which becomes an equality when the two arguments  $\bar{f}(x_0, x)$  and  $x_0(t) - x(t)$  are adjoints, i.e.

$$|x_0(t) - x(t)|^p = \text{const.} |\bar{f}(x_0, x)|^q, \quad p^{-1} + q^{-1} = 1$$

Then from (15) it follows that the best change or variation of the control signal should be adjoint to the mean gradient of performance measure.

*Remark II.* It is assumed that the weak differential  $dF(x, h)$  of  $F(x)$  is a linear functional with respect to  $h$  and therefore it can be written in the scalar product form  $dF(x, h) = [f(x), h]$ , where  $h, x \in L^p[0, T]$ ,  $f(x) \in L^q[0, T]$ .

$$^* \text{ Eqn (12):} \quad dA(x, h) = n \int_0^T k_1(t-\tau) \left[ \int_0^T k_2(\tau-\tau_1) x(\tau_1) d\tau_1 \right]^{n-1} \int_0^T k_2(\tau-\tau_1) h(\tau_1) d\tau_1 d\tau \quad (12)$$

$$^\dagger \text{ Eqn (13):} \quad dA^*(x, g') = n \int_0^T k_2(\tau-t) \left[ \int_0^T k_2(\tau-\tau_1) x(\tau_1) d\tau_1 \right]^{n-1} \int_0^T k_1(\tau_1-\tau) g'(\tau_1) d\tau_1 d\tau \quad (13)$$

To prove inequality (15) let us observe that for every number  $\alpha \in [0, 1]$  we have

$$\begin{aligned} \frac{d}{d\alpha} F[x_0 + \alpha(x - x_0)] &= dF[x_0 + \alpha(x - x_0), x - x_0] \\ &= \{f[x_0 + \alpha(x - x_0)], x - x_0\} \end{aligned}$$

Integrating this relation we obtain:

$$F(x) - F(x_0) = \int_0^1 \{d\alpha f[x_0 + \alpha(x - x_0)], (x_0 - x)\}$$

Applying the Hölder inequality we get the 'maximum principle', expressed by formula (15).

The necessary condition for a minimum of  $F(x)$  can be written in the form  $\text{grad } F(x) = \theta$ , where  $\|\theta\| = 0$ ,<sup>4</sup> and for the sufficient condition the following formula is obtained.

$$d^2 F(x, h, h) \geq \gamma(\|h\|) \|h\|$$

where  $\gamma(z)$  is a non-negative function having the property  $\lim_{z \rightarrow \infty} \gamma(z) = \infty$ .

In the case of conditional minima it must be assumed that the functionals are strongly differentiable or, what is equivalent<sup>4</sup>, that the weak differentials are continuous with respect to  $x$ . When, for instance, it is required to minimize a certain  $F_1(x)$  subject to the condition  $F_2(x) = c = \text{const.}$ , then, for the necessary condition, the following equation is obtained<sup>4</sup>.

$$\text{grad } F_1(x) = \lambda \text{grad } F_2(x)$$

where  $\lambda$  is a number and, in addition, at point  $x$  one has  $\|\text{grad } F_2(x)\| > 0$ .

In the case when the time  $T$  should be minimized subject to the constraints  $\varepsilon^{(i)}(T) = 0$ ,  $i = 0, 1, \dots, n$ , in a closed sphere of  $L^p[0, T]$  space  $\|x\| \leq R$ , the problem can be solved in two independent steps:

(1) Fix  $T$  and solve the conditional optimization problem in an open sphere of  $L^p[0, T]$ , by minimizing the functional  $F(x) = \|x\| + \sum_{i=0}^n \lambda_i \varepsilon^{(i)}(T)$ , where  $\lambda_i = \text{constant}$  multipliers determined by the constraints:  $\varepsilon^{(i)}(T) = 0$ .

(2) Assuming that the norm of the solution of (1) depends monotonically on  $T$ , the minimum  $T$  which satisfies the condition  $\|x\| \leq R$  is found.

### Optimizing Processes

When  $A(x)$  is unknown one cannot solve equation (14) and find the best  $x(t)$  in the first interval  $[0, T]$ . But it is sometimes possible to construct an optimizing process  $x_n(t)$ ,  $n = 0, 1, 2, \dots$  in the consecutive intervals  $[nT, (n+1)T]$ , which converges to the best control signal. Consider, for example, the problem of minimizing (8) which is equivalent to the solution of the equation  $y_d(t) - A(x) = 0$ , or the equivalent equation

$$x = x + \kappa [y_d(t) - A(x)] = T(x) \quad (16)$$

where  $\kappa$  is a number. This equation can be solved by the iteration

$$x_{n+1}(t) = T[x_n], \quad n = 0, 1, \dots \quad (17)$$

where  $x_0(t)$  is an arbitrary function, provided the process converges, i.e. the integral distance between  $x_{n+1}$  and  $x_n$  is smaller than the distance between  $x_n$  and  $x_{n-1}$

$$\begin{aligned} & \left\{ \int_0^T |x_{n+1}(t) - x_n(t)|^p dt \right\}^{1/p} \\ &= \left\{ \int_0^T |T(x_n) - T(x_{n-1})|^p dt \right\}^{1/p} \\ &\leq \beta \left\{ \int_0^T |x_n(t) - x_{n-1}(t)|^p dt \right\}^{1/p} \end{aligned} \quad (18)$$

where  $\beta < 1$ .

Assuming that condition (18) is satisfied for every  $x$  (which sometimes can be accomplished by choosing the proper value of  $k$ ) one can construct the sequence of functions  $x_n(t)$  by applying  $x_0(t)$  to the plant, observing  $A(x_0)$ , and computing  $x_1(t) = x_0(t) + k[y_d - A(x_0)]$ , etc. The smaller  $\beta$  is, the faster this process converges to the best  $x(t)$ . Of course, a faster converging optimizing process can be constructed if one has more information about the plant. When the plant changes slowly in time this information can be collected by observing the outputs  $y_i = A(x_i)$  for known inputs  $x_i$  and interpolating the plant operator by the polynomial operator (3), or equalizing the differentials of (3) to the differentials of the plant, determined experimentally. However, one cannot apply this approach in the case when the plant characteristics vary fast in time, because all the information collected in the past becomes obsolete in the future. Therefore it is usually better to use such iteration processes which require a minimum amount of information at every stage of optimization.

In the case when  $A(x)$  is a random operator and it is necessary to find the best  $x(t)$  with respect to the expected value, i.e., if one wants to solve the equation  $x = E\{T(x)\} = S(x)$ , use can be made of an iteration scheme similar to (17) provided  $T(x)$  satisfies certain additional conditions.

*Remark III.* Namely<sup>2</sup>, let  $\chi$  be a separable Banach space and  $T_1, T_2, \dots$  a sequence of weakly independent, weakly equally distributed continuous random operators mapping the Cartesian product  $\Omega \times \chi$  into the space  $\chi$ , i.e.,

$$\mu \left\{ \bigcap_{i=1}^n [\omega: T_i(\omega, x) \in B_i] \right\} = \prod_{i=1}^n \mu [\omega: T_i(\omega, x) \in B_i], \quad B_i \in \mathcal{B},$$

$$\mu [\omega: T_i(\omega, x) \in B] = \mu [\omega: T_k(\omega, x) \in B]$$

and satisfying the conditions:

(a) for every  $x \in \chi$  there exists the Bochner integral:

$$S(x) = \int_{\Omega} T_1(\omega, x) d\mu(\omega)$$

(b) there exists a number  $\beta < 1$  such that for every  $x_1, x_2 \in \chi$  and  $n = 1, 2, \dots$

$$\mu [\omega: \|T_n(\omega, x_1) - T_n(\omega, x_2)\| \leq \beta \|x_1 - x_2\|] = 1 \quad (19)$$

Choosing the generalized random variable  $V_1$  arbitrarily and defining the mapping  $V_{n+1}$  ( $n = 1, 2, \dots$ ) of the space  $\Omega$  into  $\chi$  for every  $\omega \in \Omega$  by the formula  $V_{n+1}(\omega) = S_n[\omega, V_n(\omega)]$ ,

where the mapping  $S_n$  of the Cartesian product  $\Omega \times \chi$  into  $\chi$  is defined by the formula

$$S_n(\omega, x) = \frac{1}{n} \sum_{i=1}^n T_i(\omega, x) \quad (20)$$

Then there exists a unique point  $\bar{x} \in \chi$ , such that  $S(\bar{x}) = \bar{x}$  and the sequence of generalized random variables converges strongly, almost surely (with probability one), to the fixed point  $\bar{x}$ .

The assumption (19) of this theorem can be relaxed, as was shown by Hanš<sup>2</sup>, by assuming

$$\mu[\omega: \|T_n(\omega, x) - T_n(\omega, \bar{x})\| \leq \beta \|x - \bar{x}\|] = 1$$

or

$$\|S(x) - \bar{x}\| \leq \beta \|x - \bar{x}\| \quad \beta < 1$$

The equation (20) can also be substituted by:

$$\bar{S}_n(\omega, x) = \frac{1}{k_n} \sum_{i=1}^{k_n} T_{j_n+i}(\omega, x)$$

where  $k_1, k_2, \dots, j_1, j_2, \dots$  are two sequences of positive integers:

$$j_1 = 1, j_{n+1} = \sum_{i=1}^n k_i + 1, n = 1, 2, \dots, \sum_{i=1}^{\infty} [1/k_n] < \infty$$

Then each realization of the process can be used only once and the control  $x$  is changed less frequently the further one proceeds.

Now check whether a similar iterative approach can be applied in the more general case of the solution of (14). Assuming that  $dA^*(x, g')$  satisfies condition (18), one can observe that this can be accomplished if we can determine experimentally the functions  $dA^*[x_n, g'(x_n)]$ , for every function  $x_n(t)$  and  $g'(x_n)$ . As an example consider the plant described by (11) for  $n = 1$  and  $k_1(t) = k_2(t) = 0$  for  $t < 0$ . The differentials (12) and (13) become linear adjoint operators, i.e.,

$$dA(x, g') = \int_0^t k(t-\tau) g'(\tau) d\tau,$$

$$dA^*(x, g') = \int_t^T k(\tau-t) g'(\tau) d\tau$$

where

$$k(t) = \int_0^t k_1(t-\tau) k_2(\tau) d\tau, \quad k(t) = 0 \text{ for } t < 0$$

Then it is easy to determine the function  $f^*(t) = dA^*[x, g'(\tau)]$  from  $f(t) = dA[x, g'(T-\tau)]$ , which can be determined experimentally by reversing in time the input  $g'(T-\tau)$  and output  $f(T-t)$ . Indeed,

$$\begin{aligned} f^*(t) &= \int_t^T k(\tau-t) g'(\tau) d\tau \\ &= \int_0^{T-t} k(T-t-\tau) g'(T-t-\tau) d\tau = f(T-t) \end{aligned} \quad (21)$$

In the case of non-linear operators, e.g. (11) for  $n > 1$ , a similar relation holds only for certain types of non-linear operators. Assuming, for example,  $k_1(t) = k_2(t) = k(t) = k(-t)$  and  $f^*(t) = dA^*[x(\tau), g'(\tau)]$ ,  $f(t) = dA[x(T-\tau), g'(T-\tau)]$  it can be proved that the gradient  $f^*(t)$  can be obtained from the

differential  $dA[x, g']$  by reversing in time the inputs and outputs; i.e.  $f^*(t) = f(T-t)$ . In the case when  $k(t)$  is symmetrical rather with respect to a certain time instant  $t = T_0$ , than  $t = 0$ , which can be regarded as the delay (or the slope of phase characteristics of the linear parts of the non-linear operator) one can find  $f^*(t)$  in an analogous way from the relation  $f^*(t) = f(T + T_0 - t)$ . The same approach can also be applied to plants described by the sum of:

(1) Linear, delayed by  $T_0$ , operator:

$$\int_0^{t-T_0} k(t-T_0-\tau) x(\tau) d\tau, \quad k(t) = 0, \quad t \leq 0$$

(2) Non-linear, delayed by  $T_0$ , operator:  $\phi[x(-T_0 + t)]$ , where  $\phi(x)$  is a non-linear function.

(3) Non-linear operator of the general type (3) which does not change when substituting  $t = T + T_0 - t$ ,  $\tau_i = T - \tau_i$ , and interchanging the integration order.

When not sure whether a particular plant belongs to the class for which the gradient can be determined by reversing inputs and outputs, one can test the required property experimentally for every input  $x(t)$ , using the following criterion:

$$\begin{aligned} &\int_0^T h_1(t) dA[x(\tau), h_2(\tau)] dt \\ &= \int_0^T h_2(t) d\bar{A}[x(T-\tau), h_1(T-\tau)] dt \end{aligned} \quad (22)$$

where  $d\bar{A}[x, h]$  denotes the reversed in time  $dA[x, h]$  operator and  $h_1(t)$ ,  $h_2(t)$  are arbitrary functions.

It can be proved that for plants which satisfy equation (22) and for which the operator  $dA[x, g']$  or  $1/\gamma \{A[x + \gamma g'] - A(x)\}$  satisfies (18) (which can be tested experimentally) the operator  $d^*A[x, g']$  (which is equal to the input-output reversed in time  $dA[x, g']$ ) will also satisfy (18), thus assuring that iteration processes of the type (17) will converge to the best  $x(t)$  when  $\gamma \rightarrow 0$  for  $n \rightarrow \infty$ .

A more general method of identification of  $dA^*[x, g']$  can be constructed using the relation

$$\int_0^T g'(t) dA[x, h] dt = \int_0^T h(t) dA^*[x, g'] dt$$

which connects  $dA[x, h]$  and  $dA^*[x, g']$ .

Indeed, the numbers

$$a_i = \frac{1}{T} \int_0^T h_i(t) dA^*[x_k, g'] dt, \quad k, i = 1, 2, \dots$$

where  $h_i(t)$  are orthogonal, i.e.

$$\begin{aligned} \frac{1}{T} \int_0^T h_i(t) h_j(t) dt &= 0, \quad i \neq j \\ &= 1, \quad i = j \end{aligned}$$

can be regarded as coefficients of the expansion of the function  $dA^*[x_k, g']$  into the series

$$dA^*[x_k, g'] = \sum_{i=1}^{\infty} a_i h_i(t)$$

Every coefficient  $a_i$  can be written as

$$a_i = \frac{1}{T} \int_0^T g'(t) dA[x_k, h_i] dt$$

$$= \frac{1}{T} \int_0^T g'(t) \lim_{\gamma \rightarrow 0} \left\{ \frac{A[x_k + \gamma h_i] - A[x_k]}{\gamma} \right\} dt$$

where  $dA[x_k, h_i]$  can be determined experimentally.

By assuming  $h_i t = T\delta(t - t_i)$  it is possible to identify  $dA^*[x_k, g']$  at any desired time moment  $t_i$ .

This method, generally speaking, requires infinite time for complete determination of  $dA^*[x_k, g']$ ,  $k = 1, 2, \dots$ . However, when the orthogonal functions  $h_i(t)$  are properly chosen a few terms of  $a_i h_i(t)$  can provide a good approximation to  $dA^*[x_k, g']$ .

When the output noise is present  $a_i$  are random variables. However, it is possible to minimize the corresponding R.M.S. error or the so-called average risk using Bayes estimates of  $a_i$ .

In this case it is also possible to improve the performance of the controller by collecting and utilizing all the past information about the plant characteristics:  $dA^*[x_k, g']$ .

It should be noted that when the gradient of the performance measure is determined experimentally many other methods, such as steepest descent or Newton generalized process, can also be constructed and applied for the plant optimization.

#### Example

For the sake of simplicity consider the non-random plant described by the operator

$$A(x) = \int_0^t k(t-\tau)x(\tau) d\tau - \varepsilon[x(t)]^2 \quad (23)$$

and the controller which minimizes the cost (7) for  $p = 2$ . The optimal iteration process corresponding to (14) is

$$x_{n+1}(t) = \frac{1}{2\lambda} dA^*[x_n(t), g'(x_n)] = \frac{1}{2\lambda} dA^*[x_n(t), 1(t)]$$

$$= \frac{1}{2\lambda} \int_t^T k(\tau-t) 1(\tau) d\tau - \frac{\varepsilon}{\lambda} x_n(t) \lambda = \lambda_1/\lambda_2 \quad (24)$$

It can be shown that the plant satisfies (22) and that the process converges if  $\beta = |\varepsilon/\lambda| < 1$ .

$$\left( d^2 F(x, h, h) = 2(\lambda_1 + \varepsilon\lambda_2) \int_0^T h^2(t) dt > 0 \text{ if } \lambda_1 + \varepsilon\lambda_2 > 0 \right)$$

Substituting  $x_0(t) = 0$  into (24) one gets  $x_1(t) = 1/2\lambda dA^*[0, 1(t)]$ . This function can be determined by applying the step function  $\gamma 1(t)$  to the plant and reversing in time the response of the plant, which is multiplied by  $(2\lambda\gamma)^{-1}$ .

For the succeeding iterations we get

$$x_n(t) = x_1(t) \left[ 1 - \frac{\varepsilon}{\lambda} + \left( \frac{\varepsilon}{\lambda} \right)^2 - \dots - \left( \frac{\varepsilon}{\lambda} \right)^{n-1} \right]$$

and

$$x(t) = \lim_{n \rightarrow \infty} x_n(t) = \frac{x_1(t)}{1 + \frac{\varepsilon}{\lambda}} \quad (25)$$

If  $\varepsilon$  were known the best  $x(t)$  could be found for the first interval by (25). When it is not known, or is changing, one can still observe the jumps  $\varepsilon/\lambda \cdot x_{n-1}(0)$ , at the beginning of every interval, and determine the value of  $\varepsilon$ . This optimizing process is shown in Figure 2 for the case  $k(t) = \alpha e^{-\alpha t}$ ,  $\alpha = 3/T$ ,  $\lambda = 1/2$ ,  $\varepsilon = 1/4$ . It is interesting to observe that the optimizing process assumes the scanning form similar to the scanning in the so-called extremum controllers. In the general case one cannot use (25) and in order to find  $x_2(t)$  should reverse in time  $1/2\lambda dA[x_1(T-\varepsilon), 1(\tau)]$ ; a procedure which is shown in Figures 3(a) and (b). One applies  $x_1(T-t)$  to the plant in the first interval and  $x_1(T-t) + \gamma 1(t)$  in the third interval; then finds [see Figure 3(b)]  $1/2\lambda\gamma \{A_I[x_1(T-\tau) + \gamma 1(t)] - A_{III}[x_1(T-\tau)]\}$  and reverses it in time to obtain  $x_2(t)$  etc. In order to utilize all the intervals we can also apply the same  $x_1(T-\tau)$  in the even intervals as shown by dotted line in Figure 3(a). When it is observed that the initial conditions of  $A(x)$  at the beginning of the adjacent intervals are the same (i.e. when the steady-state process is obtained) [see Figure 3(c)] the step signal  $\gamma 1(t)$  can be applied and the differential  $dA[x_1(T-\tau), 1(\tau)]$  can be determined.

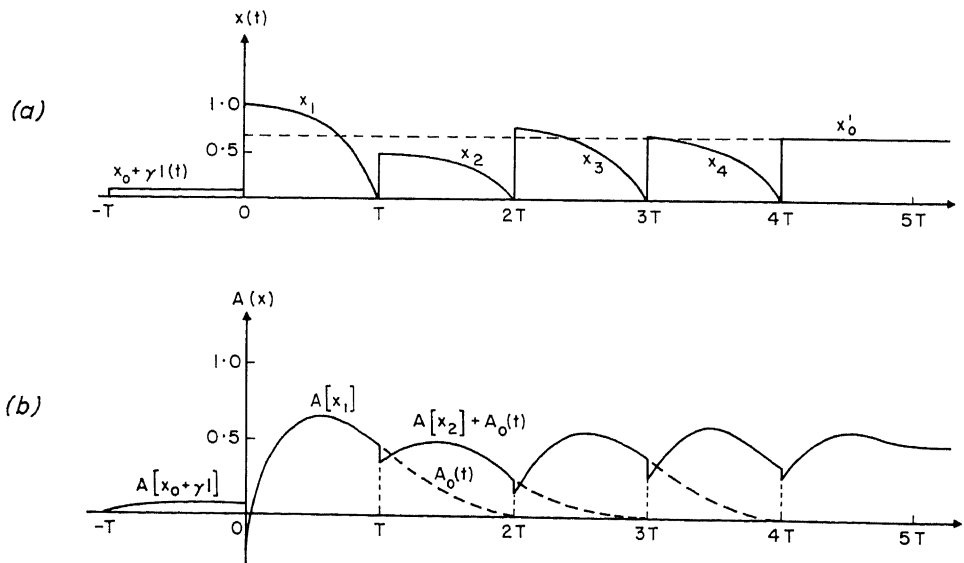


Figure 2. Construction of optimizing processes

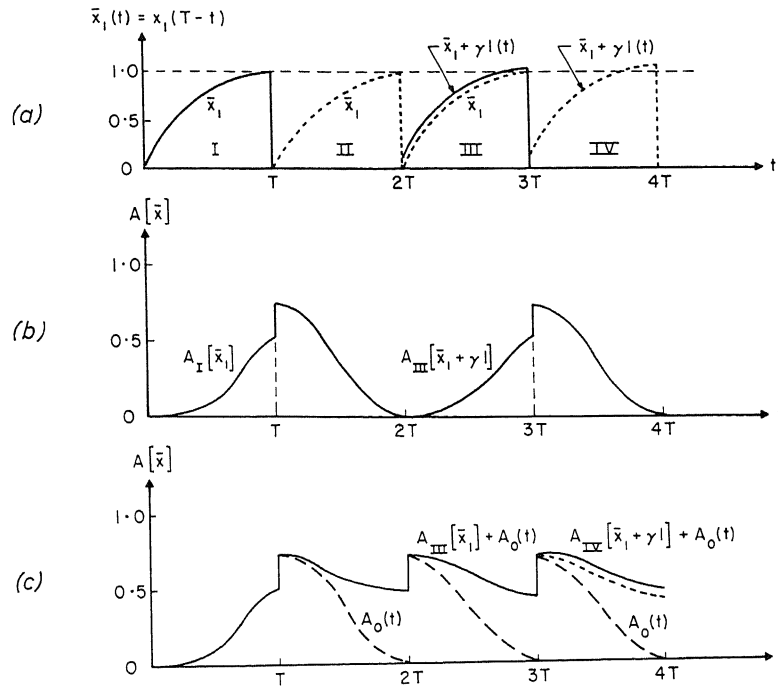


Figure 3. Construction of optimizing processes

It should be noted that the number of intervals which are necessary for the determination of the differentials can be reduced in the case when one has two identical plants or a model of the plant, e.g. two chemical reactors or two or more cylinders of the same combustion engine which are subject to the same physical conditions. In this case the differential  $dA[x_n, 1]$  for every  $x_n$  can be determined without the equalization of initial conditions. The scanning period  $T$  should be as short as possible, but not shorter than the settling time of the plant because it would not be possible to determine all the information contained in the transient process. It should be observed that in the general case of performance measure (6) the determination of the gradient is more complicated because one must reverse in time also  $g'(x_n)$  and in many cases should predict the desired state  $y_d$ .

It is interesting to compare the value of the performance measure eqn (7) for the optimum control figure eqn (25) and the corresponding value obtained when a conventional extremum-seeking regulator is being applied. Such a regulator ignores usually the input costs and transient processes in the linear part of the plant and generates an input  $x_k$  which maximizes the steady-state output of the plant. Assuming very small hunting amplitude, generated by this controller, we obtain the following best input

$$x_k = \frac{1}{2\varepsilon},$$

and the corresponding value of the performance measure is

$$F(x_k) = \frac{T}{4\varepsilon^2}(\lambda_1 - \varepsilon\lambda_2)$$

Computing the corresponding value of the performance measure  $F(x)$ , when the periodic signal eqn (25) is being applied, one gets

$$F(x) = \frac{a\lambda_2}{4\alpha}(1 - e^{-2\alpha T}) - \frac{a\lambda_2 T}{2}, \quad a = \frac{1}{2(\lambda_1 + \varepsilon\lambda_2)}$$

For a sufficiently long  $T$  (e.g.  $\alpha T > 3$ ), we obtain

$$F(x) - F(x_k) \approx -\frac{\lambda_1^2 T}{4(\lambda_1 + \varepsilon\lambda_2)\varepsilon^2} < 0$$

Because  $\lambda_1 + \varepsilon\lambda_2 > 0$ , the optimum controller which generates eqn (25) is ever better than a conventional extremum-seeking regulator.

#### References

- 1 FELDBAUM, A. A. *Computers in Automatic Systems* (in Russian). 1959. Moscow; G.I.F.M.L.
- 2 HANŠ, O. Random fixed points theorem. *Trans. Inst. Prague Conf. on Information Theory, Statistical Decision Functions, Random Processes*. Prague, 1957
- 3 KULIKOWSKI, R. On optimization of time-varying, inertial and nonlinear control systems. *Bull. Acad. Pol. Sci. Ser. Tech.* No. 8, (in Russian). 1961
- 4 WEINBERG, M. M. *Variational Methods in Nonlinear Operators* (in Russian). 1956. Moscow; G.I.T.T.L.

# Non-Linear Programming in the Investigation of Optimal Automatic Control Systems

N. I. ANDREEV

## Summary

Non-linear programming problems may be met with in the selection of optimal automatic control systems.

This paper presents a method of solving a problem in non-linear programming. The essence of the method consists in reducing the set problem to a repeated search for a solution of a linear programming problem and the choice of values for certain additional parameters that are introduced.

## Sommaire

Des problèmes de programmation non-linéaire se rencontrent dans l'étude de systèmes d'optimisation automatique. Ce rapport présente une méthode pour la solution de problèmes de programmation non-linéaire. Le principe de cette méthode consiste à ramener le problème posé à une recherche par itération de la solution d'un problème de programmation linéaire et au choix de la valeur de certains paramètres supplémentaires à introduire dans les calculs.

## Zusammenfassung

Die Synthese von Optimalwertregelungen führt auf Probleme der nichtlinearen Programmierung. Der Aufsatz enthält eine Methode zur Lösung eines solchen Problems der nichtlinearen Programmierung. Im wesentlichen besteht die Methode darin, das gegebene Problem auf ein wiederholtes Suchen nach einer Lösung eines Problems der linearen Programmierung zu reduzieren und in der Wahl der Werte von bestimmten zusätzlich eingeführten Parametern.

This paper presents a method of solving a problem in non-linear programming. The essence of the method consists in reducing the set problem to a repeated search for a solution of a linear programming problem and the choice of values for certain additional parameters that are introduced. Non-linear programming problems of a similar nature may be met with in the selection of optimal automatic control systems.

Presently linear programming has deeply penetrated into the techniques used for investigating automatic control systems. Academician Pontryagin's method<sup>1</sup>, which determines the optimal control for an automatic system in a number of practically important cases (e.g. the solution of the problem of optimal linear high-speed action), contains a linear programming problem as one of its intermediate stages. Bellman's method of dynamic programming<sup>2</sup>, which is of great generality and is also used for investigating automatic control systems, has a linear programming problem as an intermediate stage in a number of cases (when the profit function is linearly dependent on the selected parameters). Linear programming methods are used for solving reliability problems<sup>3</sup>, problems of rational tolerances in the production of assemblies<sup>4</sup>, and many other problems closely connected with the investigation and development of

automatic control systems. It should be noted that in a number of practically important cases the investigation of automatic control systems reduces to a complex problem—a non-linear programming problem, whose solution has so far only been obtained for certain particular cases<sup>4</sup>.

This paper puts forward a method of non-linear programming suitable for the solution of a broad range of problems. This method relies essentially on the techniques of linear programming. Therefore the formulation of the linear programming problem is set out below.

As is known<sup>4-6</sup>, this problem is expressed in the following manner. It is necessary to find the greatest value of a linear function of  $n$  variables  $x_1, x_2, \dots, x_n$

$$L = L(x_1, x_2, \dots, x_n) = p_1 x_1 + p_2 x_2 + \dots + p_n x_n \quad (1)$$

when the variables are subject to constraints of the form

$$\left. \begin{aligned} a_{11}x_1 + \dots + a_{n1}x_n &= b_1 \\ \dots &\dots \\ a_{1m}x_1 + \dots + a_{nm}x_n &= b_m \end{aligned} \right\} \quad (2)$$

$$\left. \begin{aligned} d_{11}x_1 + \dots + d_{n1}x_n &\leq l_1 \\ \dots &\dots \\ d_{1r}x_1 + \dots + d_{nr}x_n &\leq l_r \end{aligned} \right\} \quad (3)$$

The relations (2) and (3) determine the region  $G$  of variation of the variables  $x_1, \dots, x_n$ . These conditions can be transformed in such a way that either  $m$  or  $r$  becomes zero<sup>4</sup>. In actual problems one uses the method of writing the conditions that is most convenient.

In actual problems the function  $L$  serves as an index of the quality of the solution. The parameters  $x_1, \dots, x_n$  are characteristic of the object and the investigation, and have various physical significances according to the problem. For example, in solving a problem on high-speed action these parameters appear as control actions.

A geometrical interpretation can be given to the linear programming problem as follows: it is required to find the greatest value of linear function  $L$  of the variables  $x_1, \dots, x_n$ , whose variation is confined to a region  $G$  given in the form of a polyhedron in  $n$ -dimensional space.

Efficient techniques have been developed for solving the linear programming problem<sup>4, 5</sup>. But the linear programming method is inapplicable when either the quality index is a non-linear function  $F(x_1, \dots, x_n)$  or the region  $G$  of variation of the

parameters  $x_1, \dots, x_n$  is determined by non-linear relations between them. Such cases arise, for example, in solving the high-speed action of a system:

1. If the equations of the system include non-linear terms in the control parameters (quality index a non-linear function of the parameters);

2. If the region  $G$  of variation of the parameters is determined by non-linear relations, e.g. of the form

$$x_1^2 + \dots + x_n^2 \leq R^2$$

(the region of control forms a hypersphere centred on the origin).

If only the relations defining the region  $G$  are non-linear while the quality index is a linear function, one can replace this region by one bounded by relations of the same type as (2) and (3) which coincide accurately enough with the original region (e.g. the hypersphere may be replaced by a polyhedron circumscribed to it). The problem is thus reduced to one of linear programming.

If the quality index is a non-linear function  $F$  while the region  $G$  is determined by linear relations such as (2) and (3), one can sometimes replace the non-linear function  $F(x_1, \dots, x_n)$  by one that is piecewise linear, and proceed to solve the problem by linear programming methods<sup>4</sup>. But this device cannot always be used, and involves very bulky computation when it is applicable.

In view of what has been said, the following formulation of the non-linear programming problem is of practical and theoretical interest. Let the quality index be a given non-linear function  $F$  of the variables  $x_1, \dots, x_n$ . Without loss of generality, it may be considered that the function  $F(x_1, \dots, x_n)$  may be represented as a function  $\Phi$  of certain linear forms  $L_1, L_2, \dots, L_{k+1}$

$$F(x_1, x_2, \dots, x_n) = \Phi(L_1, L_2, \dots, L_{k+1}) \quad (4)$$

where  $\Phi$  is a given function of the variables  $L_1, \dots, L_{k+1}$ ;

$$L_i = q_{i0} + q_{i1}x_1 + \dots + q_{in}x_n \quad (5)$$

the  $q_{ij}$  being given numbers for  $i = 1, 2, \dots, n$  and  $j = 0, 1, 2, \dots, n$ , while  $k < n$ .

It is required to find the greatest value of the function  $F$  under the conditions (2) and (3).

Before proceeding with the solution of this problem, it must be explained why the function  $F$  is replaced by  $\Phi$ . The fact is that in many practical problems the number  $n$  of variables is large, and this severely complicates the process of finding a solution. Therefore it is worth while, if at all possible, to go over from the function  $F$  of many variables to the function  $\Phi$  depending on a lesser number of variables  $L_i$ . Such a transition, as will be seen from what follows, simplifies the procedure for obtaining a solution.

Two examples are given to illustrate this method of transition to a smaller number of variables.

$$\text{Example 1—} F(x_1, x_2, x_3) = x_1^2 + x_2^2 + x_3^2 + 2x_2x_3.$$

This function of three variables  $x_1, x_2$  and  $x_3$  can be expressed as a function of two other variables  $L_1$  and  $L_2$ :

$$F(x_1, x_2, x_3) = \Phi(L_1, L_2) = \frac{L_1^2 + L_2^2}{2}$$

where  $L_1 = x_1 + x_2 + x_3$ ,  $L_2 = x_1 - x_2 - x_3$ .

Here  $n = 3$  and  $k = 1$ .

$$\text{Example 2—} F(x_1, x_2) = x_1^2 + x_2^2.$$

This function of two variables cannot be expressed as a function of a lesser number of variables  $L_i$ . In this example one may put  $L_1 = x_1$  and  $L_2 = x_2$ . Here  $n = 2$  and  $k = 1$ .

The greatest value of the function  $F$  in the region  $G$  of variation of the variables  $x_1, \dots, x_n$  as defined by conditions (2) and (3) coincides with the greatest value of the function  $\Phi$  in the region  $Q$  of variation of the variables  $L_1, \dots, L_{k+1}$  as determined also in the final analysis by conditions (2) and (3). The greatest value of  $\Phi$  may be attained either within the region  $Q$  or on its boundary  $S$ . Consider each of these cases separately.

#### First Case

Suppose the function  $\Phi$  attains a maximum within the region  $Q$ . In this event the problem reduces to finding a maximum of a function of  $k + 1$  variables. It is known that a necessary condition for  $\Phi$  to have a maximum is that its partial derivatives should vanish:

$$\frac{\partial \Phi}{\partial L_i} = 0, \quad i = 1, 2, \dots, k + 1 \quad (6)$$

at a certain point in the region  $Q$  of the parameter space  $(L_1, \dots, L_{k+1})$ .

If  $\Phi$  is not differentiable everywhere inside  $Q$ , some of the conditions (6) may be replaced by these:

$$\partial \Phi / \partial L_v \text{ does not exist, } v = 1, 2, \dots, m \leq k + 1.$$

In the general case the system of eqn (6) may have several solutions. Out of them must be chosen the one that corresponds to the greatest value of  $\Phi$ . Suppose this solution has been found:

$$L_i = L_{im}, \quad i = 1, 2, \dots, k + 1 \quad (7)$$

Substituting the values (7) of the variables  $L_i$  in eqn (5), it is possible to determine the values of the quantities  $x_j = x_{jm}$  at which the required greatest value of  $F$  is attained. Thus, in this case, the problem is solved by using the normal methods of classical analysis. Conditions (2) and (3) are here used only to reject those maxima of  $\Phi$  (or  $F$ ) that do not fall within  $Q$  (or  $G$ ). This first case is rarely met in practice, since the quality index is normally taken as a function  $F$  which has no maximum within the region  $G$ . The case considered below is of greater practical interest.

#### Second Case

Suppose the function  $\Phi$  has no maximum within the region  $Q$ , and attains its greatest value on the boundary  $S$  of this region. In this case the determination of the greatest value cannot be solved by the techniques of classical analysis, and so the following two-stage method is proposed for solving this problem.

In the first stage one must determine the boundary  $S$  of the region  $Q$ , while in the second, one finds the greatest value of the function  $\Phi$  on  $S$ . Here one may make use of the ideas and techniques developed by the author<sup>7, 8</sup>, applying them to a problem of a different nature.

To determine the boundary  $S$  one may proceed in the following manner. For fixed values of the variables

$$L_1 = C_1, L_2 = C_2, \dots, L_k = C_k \quad (8)$$

one must find the greatest (and least) value of  $L_{k+1}$  (see Figure 1, where  $k = 1$ ).

Since the greatest and least values of  $L_{k+1}$  are determined by similar means, from now on only the greatest values of  $L_{k+1}$  are mentioned (i.e. only one half-branch of  $S$  is dealt with).

It follows that, to find one point on  $S$ , one must obtain the greatest value of the linear form  $L_{k+1}$  under conditions (2), (3) and (8). This is a typical linear programming problem. Conditions (8) have essentially changed nothing in conditions (2) and (3); the number of equations has merely increased by  $k$ . Taking various values of the parameters  $C_1, C_2, \dots, C_k$ , one can also derive the points on  $S$  corresponding to them. If these points are chosen so as to cover the whole of  $S$  densely enough, the first stage of the problem may be considered solved.

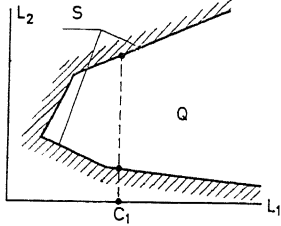


Figure 1

Now it is necessary to solve the second stage of the problem, i.e. to find the greatest value of  $\Phi$  on  $S$ . This is easily solved if the number  $k$  of dimensions of  $S$  is small. In this case the greatest value of  $\Phi$  can be determined approximately by comparing the values of  $\Phi$  at the nodes of a network formed by discrete values of the numbers  $C_1, C_2, \dots, C_k$ . If the number  $k$  of dimensions of  $S$  is large, producing a close network of values of  $\Phi$  on  $S$  becomes an extremely laborious task, which cannot always be carried out in a reasonable time even by the use of modern high-speed computer techniques.

In this case the determination of the greatest value of  $\Phi$  reduces to finding the maximum of the function

$$f = f(C_1, C_2, \dots, C_k) = \Phi[C_1, \dots, C_k, L_{k+1}(C_1, \dots, C_k)] \quad (9)$$

where  $L_{k+1}(C_1, \dots, C_k)$  is the greatest (least) value of the linear form  $L_{k+1}$  under conditions (2), (3) and (8). The greatest value of  $f$  on  $S$  in general coincides with the maximum of this function, i.e. is attained within the region of variation of the parameters  $C_i$ .

To determine the maximum of  $f = f(C_1, \dots, C_k)$  use can be made of the method of most rapid descent<sup>8,9</sup>. The combination of the method given above (which leads to the boundary  $S$  of the region  $Q$ , and to the function  $f$ ) and the method of most rapid descent (leading to the maximum of  $f$ ) makes it possible to avoid the computation of values of  $f$  at a large number of points densely covering the whole region  $S$  of variation of the variables  $C_1, \dots, C_k$ , and to replace these bulky calculations by more economic ones according to the following plan.

Let a first approximation to the variables

$$C_1 = C_{11}, C_2 = C_{21}, \dots, C_k = C_{k1}$$

be chosen from any considerations. To this corresponds a value of the function  $f_1 = f(C_{11}, C_{21}, \dots, C_{k1})$ . Now the direction of the gradient of  $f$  at this point is determined, which as is known is given by a vector in the space  $G = (C_1, \dots, C_k)$ , whose projections on the  $C_1, C_2, \dots, C_k$  axes are respectively

$$\frac{\partial f}{\partial C_1}, \frac{\partial f}{\partial C_2}, \dots, \frac{\partial f}{\partial C_k}$$

The partial derivatives  $\partial f / \partial C_i$  may be derived analytically if one has succeeded in obtaining a simple analytical expression for  $f$ .

But one cannot count on this, since normally the expression for  $f$  is complicated and, what is more, cannot be derived in explicit form. Thus in the general case the derivatives  $\partial f / \partial C_i$  must be obtained approximately as the ratio of finite differences

$$\frac{\partial f}{\partial C_i} \approx \frac{\Delta f_i}{\Delta C_i}$$

where  $\Delta f_i = f(C_{11}, \dots, C_{(i-1)1}, C_{i1} + \Delta C_i, C_{(i+1)1}, \dots, C_{k1}) - f(C_{11}, \dots, C_{k1})$ .

After determining the gradient of  $f$  at the point  $(C_{11}, \dots, C_{k1})$ , a displacement in the space  $G$  is made along this gradient vector, i.e. the values of  $f$  are considered for the following values of the variables:

$$C_1 = C_{11} + \frac{\partial f}{\partial C_1} \cdot \varepsilon, C_2 = C_{21} + \frac{\partial f}{\partial C_2} \cdot \varepsilon, \dots, C_k = C_{k1} + \frac{\partial f}{\partial C_k} \cdot \varepsilon$$

where the  $\partial f / \partial C_i$  ( $i = 1, 2, \dots, k$ ) are evaluated at  $C_1 = C_{11}, C_2 = C_{21}, \dots, C_k = C_{k1}$ .

The displacement in the chosen direction is terminated at the value  $\varepsilon = \varepsilon_1$  at which the function

$$\xi_1(\varepsilon) = f\left(C_{11} + \frac{\partial f}{\partial C_1} \varepsilon, \dots, C_{k1} + \frac{\partial f}{\partial C_k} \varepsilon\right)$$

reaches a maximum. This maximum of  $\xi_1(\varepsilon)$  may be determined graphically (see Figure 2).

The values of the variables

$$C_1 = C_{12} = C_{11} + \frac{\partial f}{\partial C_1} \varepsilon_1, \dots, C_k = C_{k2} = C_{k1} + \frac{\partial f}{\partial C_k} \varepsilon_1$$

are taken as the second approximation. The value of the function  $f = f_2 = f(C_{12}, \dots, C_{k2})$  is taken as the second approximation to  $f$ .

Then the third and succeeding approximations to the variables  $C_1, \dots, C_k$  and the function  $f$  are obtained by the method given above.

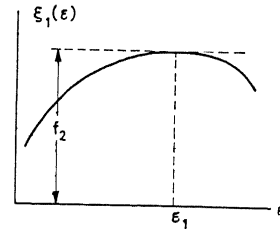


Figure 2

The  $(v+1)$ th approximation is given by

$$C_{1(v+1)} = C_{1v} + \frac{\partial f}{\partial C_1} \cdot \varepsilon_v, \dots, C_{k(v+1)} = C_{kv} + \frac{\partial f}{\partial C_k} \cdot \varepsilon_v$$

where  $\varepsilon = \varepsilon_v$  corresponds to a maximum of the function

$$\xi^v(\varepsilon) = f\left(C_{1v} + \frac{\partial f}{\partial C_1} \varepsilon, \dots, C_{kv} + \frac{\partial f}{\partial C_k} \varepsilon\right)$$

$$f_{v+1} = f(C_{1(v+1)}, C_{2(v+1)}, \dots, C_{k(v+1)})$$

The process of finding the maximum of  $f$  is terminated when two successive approximations to  $f$  differ by a negligible amount.



In the general case the function  $f$  may have several maxima, and it is necessary to find the greatest of these. It should be noted that in the general case the greatest value of  $f$  is attained within the region  $S$  of variation of the parameters  $C_1, C_2, \dots, C_k$ , i. e. it coincides with a maximum of this function. Only in rare individual cases is the greatest value of  $f$  attained on the boundary of the region  $S$ . This assertion follows from the fact that in the general case the function  $F(x_1, x_2, \dots, x_n)$  reaches its greatest value on a face, and not at a vertex, of the polyhedron defined by conditions (2) and (3).

Based on the above, the following sequence of operations can now be recommended for determining a maximum of the function  $f$ .

(a) Choose the first approximation to the variables

$$C_1 = C_{11}, \dots, C_k = C_{k1}.$$

(b) Compute the value of the first approximation to the function  $f = f_1 = f(C_{11}, \dots, C_{k1})$ .

(c) Evaluate the components of the gradient vector of  $f$  at the first approximation point:

$$\frac{\partial f}{\partial C_1}, \dots, \frac{\partial f}{\partial C_k}$$

(d) Calculate the function

$$\xi_1(\varepsilon) = f\left(C_{11} + \frac{\partial f}{\partial C_1} \varepsilon, \dots, C_{k1} + \frac{\partial f}{\partial C_k} \varepsilon\right)$$

for increasing values of the parameter  $\varepsilon = \Delta\varepsilon \cdot l$ , where  $l = 1, 2, \dots$ . The increment  $\Delta\varepsilon$  is chosen in accordance with the peculiarities of  $f$  that become evident during the process of computation: the more gentle the variation in  $f$ , the larger can  $\Delta\varepsilon$  be taken.

(e) Determine the value of the parameter  $\varepsilon = \varepsilon_1$  that makes the function  $\xi_1(\varepsilon)$  a maximum.

(f) Determine the second approximation

$$C_{12} = C_{11} + \frac{\partial f}{\partial C_1} \varepsilon_1, \dots, C_{k2} = C_{k1} + \frac{\partial f}{\partial C_k} \varepsilon_1$$

(g) Evaluate the second approximation to  $f$ :

$$f_2 = f(C_{12}, C_{22}, \dots, C_{k2}).$$

(h) Calculate the difference between the two successive approximations to  $f$ , i. e.  $f_2 - f_1$ .

This sequence is continued until the difference

$$f_{i+1} - f_i = f(C_{1(i+1)}, \dots, C_{k(i+1)}) - f(C_{1i}, \dots, C_{ki})$$

becomes negligibly small. Ordinarily the number of approximations that have to be taken when using this technique is not great. The computations involved can readily be programmed for a computer dealing with finite differences.

One may naturally wonder whether the method of most rapid descent cannot be applied directly to determining the greatest value of the function  $F(x_1, \dots, x_n)$  under conditions (2) and (3). In principle, this approach is also possible, but it leads to substantially more complex calculations in the cases where (a) the number  $n$  of variables  $x_1, \dots, x_n$  is significantly greater than the number  $k$  of variables  $C_1, \dots, C_k$ , and (b) the number of inequalities (and equations) in conditions (2) and (3) is large.

The considerable increase in the volume of computation in the first case needs no explanation. In the second case, it arises from the fact that the direct application of the method of most

rapid descent here requires that at each step of the calculation, when  $\varepsilon$  is increased by  $\Delta\varepsilon$ , one has also to check whether or not conditions (2) and (3) are satisfied. Also the transition from one face to another of the polyhedron defined by (2) and (3) involves a change in the form of a function of  $n - r$  variables.

This complicates the programming of the computation. It follows that the volume of work in deriving each approximation increases, and so does the number of approximations.

When the number of inequalities in (3) is small and  $k + 1 = n$ , both the methods become roughly equal in time-consumption.

These two different cases have been considered above: (a) The greatest value of  $F$  (and  $\Phi$ ) is attained within the region of variation of the variables  $x_1, \dots, x_n$  (or  $L_1, \dots, L_{k+1}$ ), and (b) the greatest value of  $F$  (and  $\Phi$ ) is attained on the boundary of this region, the function having no maximum within the region  $G$  (or  $Q$ ).

The case may arise [although also improbable, as (a) above] where the function  $F$  (or  $\Phi$ ) has a maximum within the region  $G$  (or  $Q$ ), but attains its greatest value on the boundary of this region. Consequently in this case the maximum of  $\Phi$  has to be found and compared with the greatest value of this function reached on the boundary  $S$ , and the greater of the two has to be chosen.

It may be expected that the techniques of solving non-linear programming problems will develop in the future, and that experience in this field will accumulate. Therefore it is worth making the following more general statement of the problem.

Let there be a method for determining the greatest (and least) value of the function  $\Psi(x_1, \dots, x_n)$  under conditions (2) and (3) that are imposed on the region of variation of the variables  $x_1, \dots, x_n$ . It is necessary to find the greatest value of the function

$$F(x_1, \dots, x_n) = \Phi(\Psi, L_1, \dots, L_k) \quad (10)$$

where  $L_p = q_{p0} + q_{p1}x_1 + \dots + q_{pn}x_n$ ,  $p = 1, 2, \dots, k$ ,  $k < n$ , and conditions (2) and (3) are satisfied.

Consider  $\Phi$  as a function of the  $k + 1$  parameters  $\Psi, L_1, \dots, L_k$ . The greatest value of this function may be attained either within the region  $Q$  of variation of these variables or on its boundary.

If the greatest value of  $\Phi$  is reached within  $Q$  (an improbable case in practice), then the problem reduces to finding the maxima of this function, which are determined by the equations:

$$\frac{\partial \Phi}{\partial \Psi} = 0 \quad \frac{\partial \Phi}{\partial L_p} = 0 \quad p = 1, 2, \dots, k \quad (11)$$

These equations enable one to determine the values of the functions  $\Psi = \Psi_0$ ,  $L_1 = L_{10}, \dots, L_k = L_{k0}$ , which correspond to a maximum of  $\Phi$ . If the solution of eqn (11) is not unique, then one must choose from all its solutions the one that corresponds to the greatest of the maxima of  $\Phi$ . From the relations

$$\Psi(x_1, \dots, x_n) = \Psi_0$$

$$q_{p0} + q_{p1}x_1 + \dots + q_{pn}x_n = L_{p0}, \quad p = 1, 2, \dots, k \quad (12)$$

one determines the values of the variables  $x_{10}, x_{20}, \dots, x_{n0}$  corresponding to the greatest value of the function  $\Phi$ . In the general case the solution of the system (12) is not unique.

If, however, the greatest value of  $\Phi$  is reached on the boundary  $S$  of the region  $Q$  (which is more likely in practical cases), then it is desirable to solve the problem as stated in two stages.

First, one must find the boundary  $S$ , and then determine the greatest value of  $\Phi$  on it. In determining the boundary  $S$ , it is necessary to take given values of the linear forms

$$L_1 = C_1, L_2 = C_2, \dots, L_k = C_k \quad (13)$$

and then determine the greatest and least values of the function  $\Psi(x_1, x_2, \dots, x_n)$  under conditions (2), (3) and (13).

It has been pointed out above that there is a method for solving this problem [the addition of (13) does not in principle alter conditions (2) and (3)]. Taking various given values of the parameters  $C_1, \dots, C_k$  one may obtain the corresponding values:

$$\Psi_1 = \Psi(C_1, C_2, \dots, C_k)$$

$$\Phi = \Phi[\Psi(C_1, \dots, C_k), C_1, \dots, C_k] = f(C_1, \dots, C_k)$$

$$x_i = x_i(C_1, C_2, \dots, C_k) \quad i = 1, 2, \dots, n$$

The second stage of the solution consists in the determination of the maximum of  $f = f(C_1, \dots, C_k)$  and the values of the variables

$$x_1 = x_{10}, x_2 = x_{20}, \dots, x_n = x_{n0}$$

corresponding to this maximum. This part of the solution is carried out in exactly the same way as for the first statement of the problem. A simple example is now given to explain the technique that has been proposed for solving the non-linear programming problem.

#### Example

Determination of the greatest value of the function

$$F(x_1, x_2, x_3) = x_1(x_2 + x_3)$$

under the conditions:

$$x_1 + 2x_2 + 3x_3 \leq 60$$

$$x_1 \geq 0, x_2 \geq 0, x_3 \geq 0$$

The given function  $F$  of the three variables  $x_1, x_2$  and  $x_3$  can be expressed as a function  $\Phi$  of two linear forms  $L_1$  and  $L_2$ :

$$L_1 = x_2 + x_3, L_2 = x_1$$

with

$$\Phi(L_1, L_2) = L_1 \cdot L_2$$

In this example  $\Phi$  is a monotone increasing function. Hence it has no maximum, and attains its greatest value on the boundary  $S$ . Following the procedure set out above, one determines the boundary  $S$  of the region  $Q$  of variation of the linear forms  $L_1$  and  $L_2$ . In this case the boundary is a certain curve (a one-dimensional domain).

In order to find  $S$ , it is necessary to take various given values of the linear form  $L_1$  and to evaluate for each the greatest

and least values of  $L_2$ . The question arises of how to choose these given values of  $L_1$ . This question is easily answered. The greatest and least values of  $L_1$  under the above conditions are readily obtained by linear programming methods and are:

$$0 \leq L_1 \leq 30$$

Taking a certain value  $L_1 = C_1$ , where  $0 \leq C_1 \leq 30$ , the greatest value of  $L_2$  is found (the least value of  $L_2$  is of no interest in this example, since  $\Phi$  is a monotone increasing function in the variable  $L_2$ ). This greatest value is easily obtained by linear programming methods (or by other means), and may be expressed in terms of  $C_1$  in the following form:

$$L_2 = 60 - 2C_1$$

The function  $\Phi$  can be expressed in terms of the parameter  $C_1$  as follows over the section of  $S$  that is being considered:

$$\Phi = C_1(60 - 2C_1)$$

It can readily be seen that the function  $\Phi$  on the boundary  $S$  attains a maximum at  $C_1 = C_{10} = 15$ .

Now it is easy to determine all the quantities of interest:

$$\Phi_0 = F_0 = 15(60 - 2 \cdot 15) = 450$$

$$x_{10} = 60 - 2 \cdot 15 = 30$$

The values of  $x_{20}$  and  $x_{30}$  are obtained from the equations:

$$x_2 + x_3 = 15$$

$$30 + 2x_2 + 3x_3 = 60 \quad (\text{see above conditions})$$

Solution of these equations gives the following values for  $x_{20}$  and  $x_{30}$ :

$$x_{20} = 15, \quad x_{30} = 0$$

#### References

1. PONTRYAGIN, L. S. et al. *The Mathematical Theory of Optimal Process*. 1961. Moscow; Fizmatgiz
2. BELLMAN, R. *Dynamic programming*. (Translated into Russian.) *Inostr. Lit.* 1960
3. SANDLER, D. Paper in the collected reports of the 5th U.S.A. Symposium, 1959
4. YUDIN, D. V., and GOL'SHTEYN, YE. G. *Problems and Methods of Linear Programming*. 1961. Sovetskoye Radio
5. GASS, S. *Linear Programming*. 1961. Moscow; Fizmatgiz
6. KRASOVSKIY, A. A., and POSPELOV, G. S. *Fundamentals of Automation and Technical Cybernetics*. 1962. Energoizdat
7. ANDREEV, N. I. A method of determining the optimum dynamic system from the criterion of extreme of a functional which is a given function of several other functionals. *Automatic and Remote Control*. 1961. London; Butterworths, Vol. 2, p. 707
8. ANDREEV, N. I. Determination of an optimal dynamic system from the criterion of a functional of partial form. *Automat. Telemekh., Moscow* 18, No. 7 (1957)
9. KANTOROVICH, L. V. Functional analysis and applied mathematics. *Usp. Mat. Nauk* 3, No. 6 (1948)

# Numerical Analysis of Non-linear Control Systems using the Fokker-Planck-Kolmogorov Equation

K. J. MERKLINGER

## Summary

The Fokker-Planck-Kolmogorov equation is a means of describing the response of non-linear control systems to Gaussian signals and disturbances. A method for numerical solution of the equation is presented here. The method is applied to first, second and third order examples of relay control systems with simple gaussian inputs. The error distribution is computed by marginal integration of the state vector distribution, and the mean squared error is used to test the advantage of linear or parabolic switching functions, and the validity of Booton's linear approximation.

## Sommaire

L'équation de Fokker-Planck-Kolmogorov est un moyen de décrire la réponse de systèmes de commande non linéaires à des signaux et perturbation gaussiens. Le rapport présente une méthode de résolution numérique de cette équation. La méthode s'applique à des exemples du premier, du deuxième et du troisième ordre de systèmes de commande à relais avec entrées gaussiennes simples. La distribution de l'erreur est calculée par intégration marginale de la distribution du vecteur d'état, et l'erreur quadratique moyenne est utilisée pour essayer l'avantage comparé des fonctions de commutation linéaires ou paraboliques, et la validité de l'approximation linéaire de Booton.

## Zusammenfassung

Mit Hilfe der Gleichung von Fokker, Planck und Kolmogorov kann man die Systemantwort nichtlinearer Regelsysteme auf Zufallssignale und Störungen mit Gaußscher Verteilung (Gaußsche Signale) beschreiben. Ein Verfahren für die numerische Lösung dieser Gleichung wird angegeben. Es wird auf Relais-Regelsysteme erster, zweiter und dritter Ordnung mit einfachen Gaußschen Eingangsgrößen angewendet. Die Fehlerverteilung wird durch Randintegration über die Verteilung des Zustandsvektors berechnet; der mittlere quadratische Fehler dient der Prüfung der Vorteile der linearen und der parabolischen Schaltfunktionen und zum Nachweis der Gültigkeit der linearen Näherung nach Booton.

## Introduction

The equation of Fokker, Planck and Kolmogorov is known by famous works in probability and physics, and by the paper of Barrett at the First IFAC Congress. The connection between the distribution of a Brownian particle and a partial differential equation was discovered by Einstein<sup>1</sup>, and developed by Smoluchowski, Fokker and Planck. More general equations for Markov processes were presented by Kolmogorov<sup>2</sup>, and applied to the study of dynamical systems by Andronov, Pontryagin and Witt<sup>3</sup>.

In automatic control, the equation has been demonstrated by Chuang and Kazda<sup>5</sup>, Pugachev<sup>6</sup>, Barrett<sup>7</sup> and Florentin<sup>8</sup>, as a means of describing the response of non-linear control systems to Gaussian signals and disturbances. Unfortunately, the equation

has been found largely intractable, except in special cases such as linear systems for which a general solution has been given by Wang and Uhlenbeck<sup>4</sup>. The first examples from Brownian motion studies were linear. Non-linear examples have been solved by Fuller<sup>9</sup>, Khazen<sup>10</sup> and Sawaragi et al.<sup>11</sup>, by piecing together elementary solutions for the linear regions of phase space, and satisfying some conditions along the boundary of each region, but it is not generally possible for elementary functions to satisfy the boundary conditions.

This paper presents a finite difference method by which approximate numerical solutions are obtained. The method is applied to first, second and third-order examples of relay control systems with simple Gaussian inputs. The accuracy of the approximate solution is tested by comparison with analytic results in two cases where they are available. The error distribution and mean squared error are obtained. Linear and parabolic switching functions are compared, and the accuracy of some results from statistical linearization is tested.

## The Fokker-Planck-Kolmogorov Equation

The canonical equation of motion of a control system with stationary Gaussian inputs is written

$$\dot{z} = f(z) + \xi \quad (1)$$

where the vector  $z$  represents the combined state of inputs and plant, and  $\xi$  is a Gaussian white noise vector independent of  $z$  and  $t$ . The components of  $z$  which represent an input  $x$ , where  $x$  is an  $n$ th order Gaussian process, may be  $x$  and its first  $n-1$  derivatives.

The Fokker-Planck-Kolmogorov equation is derived by Kolmogorov<sup>2</sup>, Barrett<sup>7</sup>, and others from the transition law for Markov processes, and the equation of motion (1). It may be written

$$\frac{\partial \omega(z, t | z_0 t_0)}{\partial t} = -\nabla \cdot J(z, t) = -\sum_i \frac{\partial}{\partial z_i} j_i(z, t) \quad (2)$$

where  $\omega(z, t | z_0 t_0)$  is the conditional probability density function of the state vector  $z$ , and  $J(z, t)$  is a vector given by

$$j_i(z, t) = f_i(z) \omega(z, t | z_0 t_0) - \frac{1}{2} \sum_k \eta_{ik} \frac{\partial}{\partial z_k} \omega(z, t | z_0 t_0) \quad (3)$$

The functions  $f_i$  are obtained from the equation of motion (1), and the constants  $\eta_{ik}$  are the cross spectral densities of the Gaussian noises  $\xi_i$  and  $\xi_k$ .

The Fokker-Planck-Kolmogorov equation describes a flow of probability in phase space. Probability particles diffuse under the influence of the Gaussian noise  $\xi$ , and drift (towards the origin if the system is stable) under the influence of the control

forces implicit in  $f$ . The equation is an expression of the continuity of the flow vector  $J$ .

The required solution is subject to the initial condition

$$\lim_{t \rightarrow t_0} \omega(z, t | z_0 t_0) = \delta(z - z_0) \quad (4)$$

and the normalizing condition

$$\int_R \dots \int \omega(z, t | z_0 t_0) dz_1 \dots dz_n = 1 \quad (5)$$

where  $R$  is the whole phase space. Of particular interest in control system design is the case where  $\omega$  converges with time to a stationary distribution independent of the initial condition, since statistical parameters of the response over a long period of time may be evaluated from the stationary distribution. In this case,

$$\lim_{t \rightarrow \infty} \frac{\partial \omega}{\partial t} = 0 \quad (6)$$

and we shall write

$$\lim_{t \rightarrow \infty} \omega(z, t | z_0 t_0) = \omega(z) \quad (7)$$

and obtain from (2) the stationary equation

$$\nabla \cdot J(z) = - \sum_i \frac{\partial}{\partial z_i} \{f_i(z) \omega(z)\} + \frac{1}{2} \sum_i \sum_k \eta_{ik} \frac{\partial^2}{\partial z_i \partial z_k} \omega(z) = 0 \quad (8)$$

### Boundary Conditions

It is apparent that eqn (8) does not exist unless the functions  $f_i$  and  $\omega$  are differentiable with respect to the  $z$  variables. In the case of relay control systems, the drift vector  $f$  is not differentiable on the switching surface  $C$ . Boundary conditions are then applied along  $C$  which effectively piece together solutions for the two regions bordering upon  $C$ , in which  $f$  and  $\omega$  are assumed to possess sufficient derivatives. Suitable boundary conditions are

- (a)  $J \cdot n$  is continuous, where  $n$  is a unit vector normal to  $C$ , and
- (b)  $\omega$  is continuous.

The necessity of the first condition is made apparent by the application of Gauss' Theorem to a thin volume  $V$ , parallel with and containing  $C$  (Figure 1). We can write

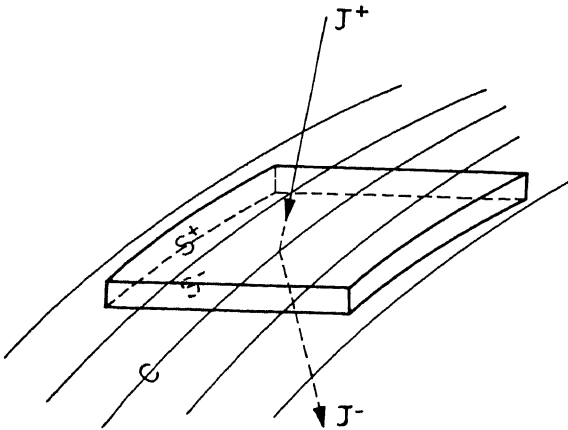


Figure 1. A thin volume containing the switching surface  $C$

$$\int_S J \cdot n dS = \int_V \dots \int \nabla \cdot J dV = - \int_V \dots \int \frac{\partial \omega}{\partial t} dV \quad (9)$$

where  $S$  is the surface of  $V$ , and  $n$  is a unit vector normal to  $S$ . Since in the stationary case,  $\partial \omega / \partial t = 0$ , one obtains

$$\int_S J \cdot n dS = 0 \quad (10)$$

In the presence of the diffusing vector  $\xi$ , it is reasonable to postulate that no singularity of particles exists on surface  $C$ . Then, if  $V$  is infinitesimally thin, there is no measurable flow of particles through the sides of  $V$ , and one can write

$$\int_S J \cdot n dS = \int_{S^+} J \cdot n dS + \int_{S^-} J \cdot n dS = 0 \quad (11)$$

where  $S^+$  and  $S^-$  are the two faces of  $V$ . In the limit, when  $S^+$  and  $S^-$  coincide, one obtains

$$J^+ \cdot n_c = J^- \cdot n_c$$

where  $J^+$  and  $J^-$  are flow vectors on opposite sides of  $C$ , and  $n_c$  is a unit vector normal to  $C$ .

The second boundary condition ( $\omega$  is continuous) is postulated on the grounds that any discontinuity in the density of particles will be dissolved as quickly as it occurs by the diffusive force of  $\xi$ .

### Finite Difference Methods

Characteristically, finite difference methods divide the domain  $R$  of the  $z$  variables into rectangular cells. The quantization may be accomplished physically or mathematically. In the physical interpretation, eqn (8) is replaced by laws for the transition of particles between cells, given the transition probabilities of a vector Markov chain with cells as states. Applied to all cells, the transition laws generate a system of difference equations. The same system of equations may be obtained by replacing the derivatives in eqn (8) with difference formulae analogous to the transition laws.

The following discussion will be restricted for simplicity, to the two-dimensional space  $R_2 = (z_1, z_2)$ . Consider a network of lines to divide  $R_2$  into rectangles, and let  $W(z_1, z_2)$  denote a function in  $R_2$ . The value of  $W$  at the nodes of the network will be denoted by

$$W(z_1, z_2) = W(mh, nk) = W_{mn}$$

where  $m$  and  $n$  are the integers  $0, \pm 1, \pm 2, \dots$ , and  $h$  and  $k$  are the dimensions of the rectangles. The nodal function  $W_{mn}$  will satisfy a difference equation

$$L_{hk}(W_{mn}) = 0 \quad (12)$$

with the desired convergence property:

$$\lim_{h, k \rightarrow 0} L_{hk}(W_{mn}) = L(\omega) = - \sum_i \frac{\partial}{\partial z_i} f_i(z) \omega(z) + \frac{1}{2} \sum_i \sum_k \eta_{ik} \frac{\partial^2}{\partial z_i \partial z_k} \omega(z) \quad (13)$$

It is often hoped that this property is sufficient to provide the further convergence,

$$\lim_{h, k \rightarrow 0} W = \omega \quad (14)$$

That it is not sufficient as an important negative statement from the finite difference theory (given, for example, by Forsythe and Wasow<sup>12</sup>).

An operator  $L_{hk}$  will be required which possesses the convergence property (13) and which, in addition, yields a solution for  $L_{hk}(W_{mn}) = 0$ , such that

$$|\omega(z_1, z_2) - W(z_1, z_2)| \quad \text{is sufficiently small} \quad (15)$$

where  $W_{mn}$  has been extended by interpolation to cover  $R_2$ . Further, one would like to obtain  $L_{hk}$  and solve (12) in the most economical way.

Unfortunately, it can rarely be determined from theory whether (15) is satisfied, and one must adopt an experimental approach such as testing a difference procedure for some problems whose exact solutions are known, or testing the procedure for various values of  $h$  and  $k$  and observing the stability of  $W$ .

### A Finite Difference Model

#### The Crank-Nicolson Formula

Consider the control systems having only one input, for which the stationary equation in two dimensions may be written

$$\frac{\partial^2 \omega(z_1, z_2)}{\partial z_1^2} + a(z_1, z_2) \frac{\partial \omega(z_1, z_2)}{\partial z_1} + b(z_1, z_2) \frac{\partial \omega(z_1, z_2)}{\partial z_2} + c(z_1, z_2) \omega(z_1, z_2) = 0 \quad (16)$$

The derivatives in the above equation will be replaced by the Crank-Nicolson difference approximation<sup>13</sup> as follows:

$$\begin{aligned} \frac{\partial^2 \omega}{\partial z_1^2} &\rightarrow \frac{1}{2h^2} \{W_{m+1, n} - 2W_{m, n} + W_{m-1, n} \\ &\quad + W_{m+1, n-u} - 2W_{m, n-u} + W_{m-1, n-u}\} \\ \frac{\partial \omega}{\partial z_1} &\rightarrow \frac{1}{4h} \{W_{m+1, n} - W_{m-1, n} + W_{m+1, n-u} - W_{m-1, n-u}\} \\ \frac{\partial \omega}{\partial z_2} &\rightarrow \frac{u}{k} \{W_{m, n} - W_{m, n-u}\} \\ \omega &\rightarrow \frac{1}{2} \{W_{m, n} + W_{m, n-u}\} \end{aligned} \quad (17)$$

where  $u = \text{sgn } C(z_1, z_2)$   
and  $C(z_1, z_2) = 0$  is the switching curve.

The arguments of the coefficients  $a, b, c$  will be interpreted as  $(mh, (n - u/2)k)$ .

Experiments with eqn (16) have shown that the Crank-Nicolson formula is more accurate than simpler asymmetric formulae, and the symmetric formula of Richardson<sup>12</sup> is unstable.

From the Crank-Nicolson formula we obtain the system of difference equations given by

$$\begin{aligned} &\left\{ \frac{1}{2h^2} + \frac{a_{m, n-u/2}}{4h} \right\} W_{m+1, n} + \left\{ \frac{1}{2h^2} + \frac{a_{m, n-u/2}}{4h} \right\} W_{m+1, n-u} \\ &+ \left\{ -\frac{1}{h^2} + \frac{u}{k} b_{m, n-u/2} + \frac{1}{2} c_{m, n-u/2} \right\} W_{m, n} \\ &+ \left\{ -\frac{1}{h^2} - \frac{u}{k} b_{m, n-u/2} + \frac{1}{2} c_{m, n-u/2} \right\} W_{m, n-u} \end{aligned} \quad (18)$$

$$+ \left\{ \frac{1}{2h^2} - \frac{a_{m, n-u/2}}{4h} \right\} W_{m-1, n} + \left\{ \frac{1}{2h^2} - \frac{a_{m, n-u/2}}{4h} \right\} W_{m-1, n-u}$$

$$= 0; \quad m, n = 0, \pm 1, \pm 2, \dots$$

### Non-uniform Cell Sizes

Since the required function  $W$  is expected to be smoothest away from the origin (if the control system is good), it will be expedient, for economy of points, to introduce the transformation

$$(z_1, z_2) \rightarrow (z'_1, z'_2) \quad (19)$$

where

$$z_i \rightarrow e^{z'_i} - 1, \quad z_i \geq 0$$

and

$$z_i \rightarrow 1 - e^{z'_i}, \quad z_i \leq 0; \quad i = 1, 2$$

The transformation (19) is equivalent to dividing the  $(z_1, z_2)$  space into non-uniform cells such that the nodes  $(m, n)$  are denser towards the origin. The system of difference equations for  $W(z'_1, z'_2) = W(mh, nk) = W_{mn}$  may be written

$$\begin{aligned} &\left\{ \frac{1}{2h^2} e^{-2mh} - \frac{1}{4h} \left[ e^{-2mh} - e^{-mh} a'_{m, n-u/2} \right] \right\} W_{m+1, n} \\ &+ \left\{ \frac{1}{2h^2} e^{-2mh} - \frac{1}{4h} \left[ e^{-2mh} - e^{-mh} a'_{m, n-u/2} \right] \right\} W_{m+1, n+u} \\ &+ \left\{ -\frac{1}{h^2} e^{-2mh} + \frac{u}{k} e^{-nk} b'_{m, n-u/2} + \frac{1}{2} c'_{m, n-u/2} \right\} W_{m, n} \\ &+ \left\{ -\frac{1}{h^2} e^{-2mh} - \frac{u}{k} e^{-nk} b'_{m, n-u/2} + \frac{1}{2} c'_{m, n-u/2} \right\} W_{m, n-u} \\ &+ \left\{ \frac{1}{2h^2} e^{-2mh} + \frac{1}{4h} \left[ e^{-2mh} - e^{-mh} a'_{m, n-u/2} \right] \right\} W_{m-1, n} \\ &+ \left\{ \frac{1}{2h^2} e^{-2mh} + \frac{1}{4h} \left[ e^{-2mh} - e^{-mh} a'_{m, n-u/2} \right] \right\} W_{m-1, n-u} \end{aligned} \quad (20)$$

$$= 0; \quad m, n = 0, +1, +2, \dots$$

and

$$\begin{aligned}
& \left\{ \frac{1}{2h^2} e^{+2mh} + \frac{1}{4h} \left[ e^{+2mh} + e^{mh} a_{m,n-\frac{u}{2}} \right] \right\} W_{m+1,n} \\
& + \left\{ \frac{1}{2h^2} e^{2mh} + \frac{1}{4h} \left[ e^{2mh} + e^{mh} a_{m,n-\frac{u}{2}} \right] \right\} W_{m+1,n-u} \\
& + \left\{ -\frac{1}{h^2} e^{2mh} + \frac{u}{k} e^{-nk} b_{m,n-\frac{u}{2}} + \frac{1}{2} c_{m,n-\frac{u}{2}} \right\} W_{m,n} \\
& + \left\{ -\frac{1}{h^2} e^{2mh} - \frac{u}{k} e^{-nk} b_{m,n-\frac{u}{2}} + \frac{1}{2} c_{m,n-\frac{u}{2}} \right\} W_{m,n-u} \\
& + \left\{ \frac{1}{2h^2} e^{2mh} - \frac{1}{4h} \left[ e^{2mh} + e^{mh} a_{m,n-\frac{u}{2}} \right] \right\} W_{m-1,n} \\
& + \left\{ \frac{1}{2h^2} e^{2mh} - \frac{1}{4h} \left[ e^{2mh} + e^{mh} a_{m,n-\frac{u}{2}} \right] \right\} W_{m-1,n-u} \\
& = 0; \quad m=0, -1, -2, \dots, \quad n=0, +1, +2, \dots, \text{ etc.}
\end{aligned}$$

### Boundary Conditions at the Switching Curve

It has been stated, in the analytic discussion, that eqn (8) does not exist on the switching curve  $C(z_1, z_2) = 0$ , and that boundary conditions are applied along  $C$  which effectively piece together solutions for the regions  $C(z_1, z_2) > 0$  and  $C(z_1, z_2) < 0$ . Similarly, the difference eqn (20) may not be written when points  $(m+1, n)$  and  $(m-1, n)$  are astride the switching curve, and difference equations derived from the analytic boundary conditions are applied here.

The boundary condition,  $\omega$  is continuous, may be applied simply. The condition  $J \cdot n$  is continuous, may be written

$$j_{z_1} \cos \theta - j_{z_2} \sin \theta \text{ is continuous} \quad (21)$$

where  $\theta$  is the angle between the  $z_2$  axis and the switching curve at  $(m', n)$ , and  $m'$  is the generally non-integral value such that

$$C(m'h, nk) = 0$$

Nodal points  $(m, n)$  will be eliminated in favour of points  $(m', n)$  on the switching curve whenever  $|m' - m| < \frac{1}{2}$  or  $m' - m = \frac{1}{2}$ . Formulae (17) must then be replaced by the corresponding formulae for unequal differencing intervals (Forsythe and Wasow<sup>12</sup>) in the case of points  $(m, n)$  such that  $|m' - m| < \frac{3}{2}$  or  $m' - m = \frac{3}{2}$ .

From eqn (16) the components of the flow vector can be written as

$$\begin{aligned}
j_{z_1} &= -\frac{\partial \omega}{\partial z_1} - a(z_1, z_2) \omega \\
j_{z_2} &= -b(z_1, z_2) \omega
\end{aligned} \quad (22)$$

and the condition (21) as

$$\left( \frac{\partial \omega}{\partial z_1} + a\omega \right) \cos \theta - b\omega \sin \theta \text{ is continuous} \quad (23)$$

or in the  $(z_1', z_2')$  space defined by (19),

$$\left( e^{-z_1'} \frac{\partial \omega}{\partial z_1'} + a'\omega \right) \cos \theta - b'\omega \sin \theta \text{ is continuous} \quad (24)$$

where  $z_1' \geq 0$ ; etc.

The derivative  $\partial \omega / \partial z_1'$ , is replaced as follows:

$$\frac{\partial \omega}{\partial z_1'} \rightarrow \frac{1}{h} \left\{ -\frac{3}{2} W_{m'n} + 2 W_{m+1,n} - \frac{1}{2} W_{m+2,n} \right\}, \quad C(z_1', z_2') = 0^+$$

$$\frac{\partial \omega}{\partial z_1'} \rightarrow \frac{1}{h} \left\{ \frac{3}{2} W_{m'n} - 2 W_{m-1,n} + \frac{1}{2} W_{m-2,n} \right\}, \quad C(z_1', z_2') = 0^-$$

converting the continuity condition (24) to a difference equation

$$\begin{aligned}
& -\frac{1}{2} W_{m-2,n} + 2 W_{m-1,n} \\
& - \{ 3 + \gamma h e^{mh} [-a^+ + a^- + (b^+ - b^-) \tan \theta] \} W_{m'n} \\
& + 2 W_{m+1,n} - \frac{1}{2} W_{m+2,n} = 0
\end{aligned} \quad (26)$$

for points  $(m', n)$  on the switching curve.

Parameter  $\gamma$  in (26) is at present gratuitous.

### Boundary Conditions on the Perimeter

Another approximation will have to be made before the system of difference equations (20) and (26) can be solved. Since the  $z$  space is unbounded, the system of equations is unbounded unless  $m$  and  $n$  are confined within limits such as

$$\begin{aligned}
|m| &\leq r \\
|n| &\leq s
\end{aligned} \quad (27)$$

It then becomes necessary to impose a boundary condition on the perimeter  $S$  defined as points  $(m, n)$  such that  $m = \pm r$ , or  $n = \pm s$ .

One may attempt to set

$$W_{mn} = 0, \quad (m, n) \in S \quad (28)$$

because of the prospect that for stable systems,

$$\lim_{|z_1|, |z_2| \rightarrow \infty} \omega(z_1, z_2) = 0$$

However, (28) may not be a good approximation since in a physical sense, it implies that  $S$  is an absorbing boundary. Alternatively, one could impose a reflecting boundary at  $S$ , but this too is unlike the analytic process where particles flow in both directions across  $S$ . A compromising condition will be adopted to simulate more accurately the analytic process.

Let

$$\begin{aligned}
W_{r+1,n} &= \alpha W_{r,n} \\
W_{-r-1,n} &= \alpha W_{-r,n} \\
W_{m,s+1} &= \beta W_{m,s} \\
W_{m,-s-1} &= \beta W_{m,-s}
\end{aligned} \quad (29)$$

where  $\alpha$  and  $\beta$  are positive constants less than unity. The outward flow by diffusion at the boundary  $m = r$  is then proportional to  $1 - \alpha$ , and may be zero ( $\alpha = 1$ ) or maximum ( $\alpha = 0$ ) similar to the extreme conditions considered above. The best values of  $\alpha$  and  $\beta$  are obtained by experiment of which more will be said.

### The Normalizing Condition

The normalizing condition (5) may be satisfied as follows. The condition may be rewritten for a finite boundary  $S$  as

$$\int_{-s}^s \int_{-r}^r W(z_1, z_2) dz_1 dz_2 = 1 \quad (30)$$

Set

$$KW_{00} = 1 \quad (31)$$

and solve for the relative values

$$KW_{mn}; (m, n) \neq (0, 0)$$

and then determine the normalizing constant  $K$  from the integration (30).

### Empirical Determination of the Parameters $\alpha, \beta, \gamma$

Setting  $KW_{00} = 1$ , as in (31), necessitates the elimination of one equation from the system of eqns (20) and (26). If the matrix of the system is not degenerate\* so that no equation is redundant, the elimination of the equation  $m, n = 0$  gives the point

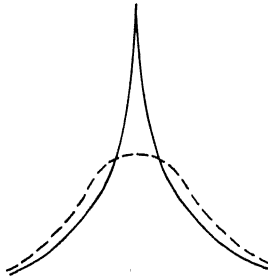


Figure 2. Cross section  $W(C)$  displaying the possible source action of the point  $(0, 0)$ ; --- denotes the correct solution

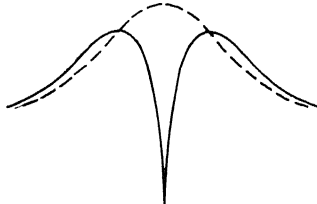


Figure 3. Cross section  $W(C)$  displaying the possible sink action of the point  $(0, 0)$ ; --- denotes the correct solution

$(0, 0)$  freedom to act as a source or sink of particles. If the values of  $\alpha$  and  $\beta$  are too low so that there is a net outflow of particles at the boundary  $S$ , the solution will have the character depicted in Figure 2, indicating that the point  $(0, 0)$  is acting as a source to supply the particles lost at  $S$ . If the values of  $\alpha$  and  $\beta$  are too high, there is a net inflow of particles (by drift) at  $S$ , and the solution will have the character depicted in Figure 3 with the point  $(0, 0)$  acting as a sink.

\* Since the coefficients of eqns (26) and (29) (and (20) if  $c \neq 0$ ) do not sum to zero, the matrix row sums cannot all be zero; because of the transformation (19) none of the column sums is zero; neither is there a sequence of numbers  $\lambda_i, i = 1, 2, \dots$  such that if the  $i$ th row (column) is multiplied by  $\lambda_i$  for  $i = 1, 2, \dots$ , the column (row) sums are zero; therefore the matrix is not degenerate and there may be a sink or source at  $(0, 0)$ .

Since the analytic solution  $\omega(z_1, z_2)$  is known to be smooth along the switching curve, the parameters  $\alpha$  and  $\beta$  may be chosen so that the numerical solution  $W(z_1, z_2), C(z_1, z_2) = 0$ , is also smooth: so that the point  $(0, 0)$  does not act as a source or sink.

Because of the nature of eqn (26), the points  $(m', n)$  may also act as sources or sinks. Small variations of the parameter  $\gamma$  ( $\gamma \leq 1$ ) enable adjustment of  $W(z_1, z_2)$  until the necessary flow continuity is achieved without injection or absorption of particles at the points  $(m', n)$ . The criterion of adjustment is again the smoothness of  $W(z_1, z_2)$  for  $C(z_1, z_2) = 0$ : the point  $(0, 0)$  must not act as a source or sink.

There is some indeterminacy in setting three parameters  $\alpha, \beta, \gamma$  from one criterion. However, it is helpful to know that

$$\lim_{r, s \rightarrow \infty} \alpha, \beta = 0$$

and

$$\lim_{h, k \rightarrow 0} \gamma = 1$$

Also, the most effective parameter is  $\gamma$  and  $\alpha$  and  $\beta$  may be used only for fine adjustments if required. It has been found in practice from examples for which analytic results are known that any set of parameters  $\alpha, \beta, \gamma$  which clearly satisfies the condition at  $(0, 0)$  produces a good approximation  $W(z_1, z_2)$  in the sense of (15).

### Computation of the Solution

The system of difference equations may be written as the linear equation

$$AX = B \quad (32)$$

where  $A$  is a matrix, and  $X$  and  $B$  are vectors for the solution and right-hand sides. Eqn (32) has been solved on a digital computer by Gaussian elimination.  $A$  is a banded matrix and only elements inside the band need to be stored. Systems of up to 900 eqns have been solved for this problem, although it will be seen that fewer equations are required for useful analysis of second and third-order control systems.

### Examples of Numerical Solutions of the Fokker-Planck-Kolmogorov Equation

#### A First-order System

Equations of motion of the relay control system depicted in Figure 4 are

$$\begin{aligned} \dot{e} &= -\frac{1}{\tau}(e + y) - \frac{1}{T} \operatorname{sgn} C + \frac{\xi}{\tau} \\ \dot{y} &= \frac{1}{T} \operatorname{sgn} C \end{aligned} \quad (33)$$

where  $C(e, y)$  is the switching function, and the input  $x$  has the spectrum  $N^2/(1 + \omega^2 \tau^2)$ .

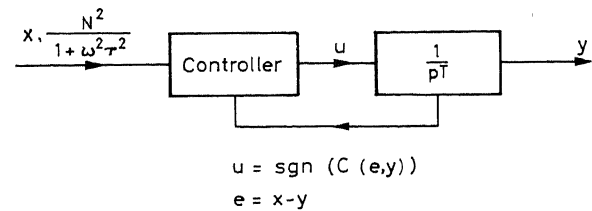


Figure 4

The Fokker-Planck-Kolmogorov equation and boundary condition for  $C(e, y) = 0$  may be written

$$\frac{N^2}{2\tau^2} \frac{\partial^2 \omega}{\partial e^2} + \left\{ \frac{1}{\tau} (e + y) + \frac{1}{T} \operatorname{sgn} C \right\} \frac{\partial \omega}{\partial e} - \frac{1}{T} \operatorname{sgn} C \frac{\partial \omega}{\partial y} + \frac{1}{\tau} \omega = 0 \quad (34)$$

and

$$\frac{N^2}{2\tau^2} \frac{\partial \omega}{\partial e} + \left\{ \frac{1}{\tau} (e + y) + \frac{1}{T} \operatorname{sgn} C \right\} \omega + \frac{g}{T} \operatorname{sgn} C \omega \text{ is continuous,}$$

where  $g = de/dy$ ,  $C(e, y) = 0$ .

Eqn (34) has been solved numerically for the case  $N^2 = 2$ ,  $T = \tau = 1$  with the linear switching function,

$$C(e, y) = e - Ky$$

The error distribution obtained by numerical integration of the  $(e, y)$  distribution is shown in Figure 5. The mean squared error is given in Table 1, with some details of the numerical solution.

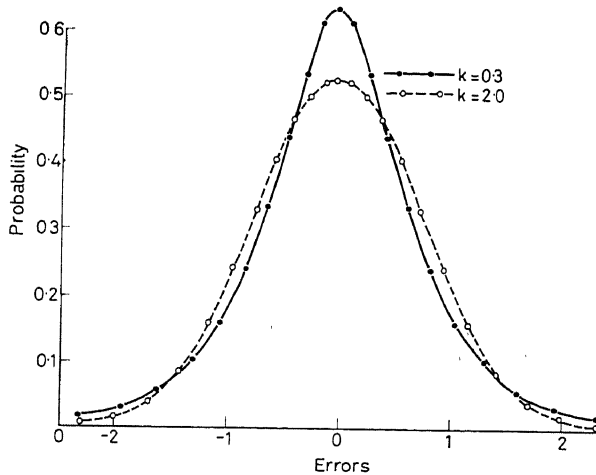


Figure 5. Error distributions

Table 1. Mean Squared Error and Parameters of the Numerical Solution for the Simple Integrator

Control parameter $K$	Mean squared error	Cell sizes		Grid dimensions		Number of equations
		$h$	$k$	$r$	$s$	
0.3	0.53514	0.12	0.12	10	8	179
0.6	0.47416	0.12	0.12	10	8	179
1.0	0.46258	0.1	0.1	12	9	238
1.4	0.47940	0.1	0.09	12	9	238
2.0	0.52198	0.11	0.069	12	10	263

The optimum value of  $K$  is near 1.0, and the minimum mean squared error is 0.46.

The accuracy of the computed  $(e, y)$  distribution has been tested by marginal integration of the  $(e + y, y) = (x, y)$  distribution, for the theoretically known  $x$  distribution. The numerical and theoretical results are in excellent agreement (Figure 6).

The advantage of parabolic switching:  $C(e, y) = e - Ky|y|$  has been investigated with the results shown in Figure 7. The minimum mean squared error is reduced by 2 per cent.

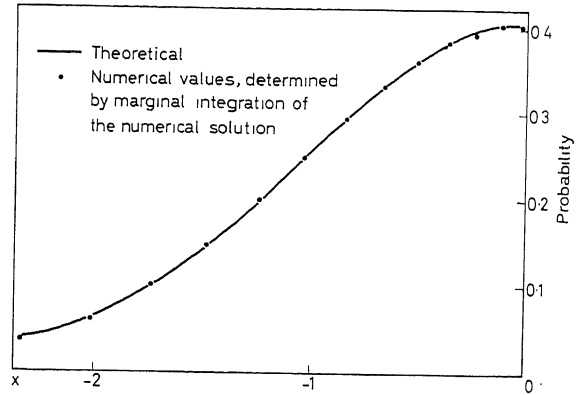


Figure 6.  $x$ -distribution

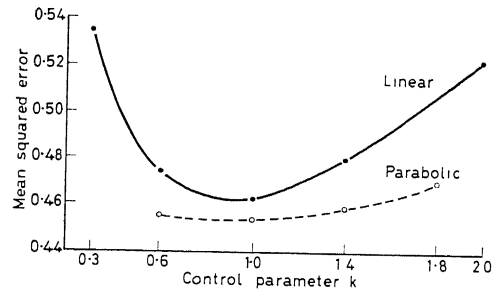


Figure 7. Linear versus parabolic switching

#### A Second-order System

As a second example, a numerical solution will be compared with the analytic solution obtained by Fuller<sup>7</sup> for the system depicted in Figure 8. The equation to be solved is

$$\frac{N^2}{2} \frac{\partial^2 \omega}{\partial e^2} + z \frac{\partial \omega}{\partial e} - \frac{1}{T} \operatorname{sgn} C \frac{\partial \omega}{\partial z} = 0 \quad (35)$$

where  $z = \dot{y}$  and  $e = x - y$ .

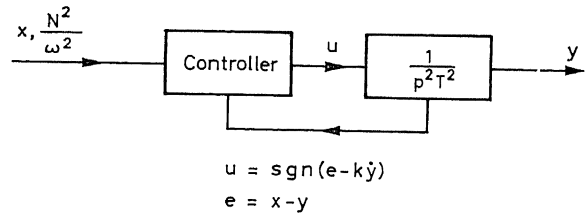


Figure 8

The analytic solution is given by

$$\omega = \frac{K^{3/2}}{\sqrt{\pi} N^3 T^2} \exp \left\{ -\frac{K}{N^2} \left( z^2 + \frac{2}{T^2} |e - Kz| \right) \right\} \quad (36)$$

and the error distribution by

$$\omega(e) = \frac{K}{2 N^2 T^2} \exp \frac{K^3}{N^2 T^2} \left\{ \exp \frac{2 K e}{N^2 T^2} \left[ 1 - \operatorname{erf} \left( \frac{e}{N \sqrt{K}} + \frac{K^{3/2}}{N T^2} \right) \right] + \exp \frac{-2 K e}{N^2 T^2} \left[ 1 + \operatorname{erf} \left( \frac{e}{N \sqrt{K}} - \frac{K^{3/2}}{N T^2} \right) \right] \right\} \quad (37)$$



Table 2. Numerical Solution for the Double Integrator  $K = 0.2$  with Theoretical Values in Parentheses. The  $(e, \dot{y})$  Distribution is Symmetrical with Respect to the Origin

The $(e, \dot{y})$ distribution:						
$\uparrow \dot{y}$						
0.8158 (0.8194)	0.8568 (0.8605)	0.8138 (0.8191)	0.7605 (0.7668)	0.7004 (0.7073)	0.6332 (0.6409)	0.5597 (0.5683)
0.8971 (0.9044)	0.9280 (0.9347)	0.8698 (0.8754)	0.8136 (0.8195)	0.7491 (0.7559)	0.6769 (0.6850)	0.5981 (0.6073)
0.9739 (0.9761)	0.9496 (0.9524)	0.8985 (0.9022)	0.8394 (0.8446)	0.7722 (0.7791)	0.6974 (0.7060)	0.6159 (0.6259)
0.9937 (0.9951)	0.9585 (0.9605)	0.9064 (0.9099)	0.8464 (0.8517)	0.7784 (0.7857)	0.7027 (0.7120)	0.6205 (0.6312)
1.0000 (1.0000)	0.9546 (0.9567)	0.9024 (0.9063)	0.8424 (0.8484)	0.7745 (0.7826)	0.6991 (0.7092)	0.6173 (0.6288)
						$\rightarrow (e)$
	0.9413 (0.9436)	0.8895 (0.8939)	0.8300 (0.8368)	0.7632 (0.7719)	0.6890 (0.6995)	0.6083 (0.6202)
	0.9110 (0.9156)	0.8610 (0.8674)	0.8041 (0.8120)	0.7397 (0.7490)	0.6677 (0.6787)	0.5892 (0.6018)
	0.8585 (0.8652)	0.8127 (0.8197)	0.7592 (0.7673)	0.6980 (0.7078)	0.6295 (0.6414)	0.5550 (0.5687)
	0.7785 (0.7839)	0.7353 (0.7426)	0.6857 (0.6951)	0.6296 (0.6413)	0.5673 (0.5811)	0.4998 (0.5152)
The normalizing constant: 8.5734 (8.6630)						
The error distribution:						
0.2661 (0.2645)	0.2595 (0.2580)	0.2453 (0.2443)	0.2291 (0.2286)	0.2107 (0.2109)	0.1903 (0.1911)	0.1679 (0.1695)
The mean squared error: 1.5840 (1.5909)						

The mean squared error is

$$\frac{N^2}{2} \left( \frac{N^2 T^4}{k^2} + K \right) \quad (38)$$

Tables 2, 3 and 4 present the numerical solution and theoretical values for  $K = 0.2, 1.0, 1.8$ , together with numerical and theoretical values for the error distribution and mean squared error.

The numerical solutions are obtained for a relatively small number of points, 59 or 68, to demonstrate the accuracy which can be achieved by a small system of equations. The estimate of mean squared error is in error by 0.4 to 1.3 per cent.

### A Third-order System

A solution at 83 points for the system shown in Figure 9 is given in Table 5.

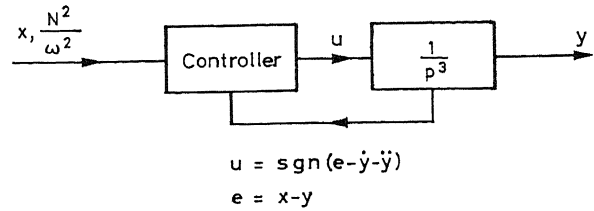


Figure 9

Table 3. Numerical Solution for the Double Integrator  $K = 1.0$  with Theoretical Values in Parentheses. The  $(e, \dot{y})$  Distribution is Symmetrical with Respect to the Origin

The $(e, \dot{y})$ distribution:						
$\uparrow \dot{y}$						
0.1518 (0.1386)	0.1867 (0.1729)	0.2397 (0.2266)	0.3242 (0.3153)	0.4673 (0.4719)	0.2843 (0.2883)	0.1560 (0.1579)
0.3250 (0.3135)	0.3992 (0.3912)	0.5117 (0.5126)	0.6909 (0.7132)	0.4646 (0.4765)	0.2872 (0.2911)	0.1592 (0.1595)
0.5343 (0.5419)	0.6575 (0.6761)	0.8462 (0.8861)	0.6178 (0.6368)	0.4191 (0.4254)	0.2599 (0.2599)	0.1446 (0.1424)
0.7600 (0.7820)	0.9445 (0.9758)	0.7301 (0.7446)	0.5292 (0.5351)	0.3570 (0.3575)	0.2215 (0.2184)	0.1238 (0.1197)
1.0000 (1.0000)	0.7953 (0.8014)	0.6053 (0.6115)	0.4369 (0.4395)	0.2957 (0.2936)	0.1847 (0.1794)	0.1039 (0.0983)
						$\rightarrow (e)$
	0.6129 (0.6267)	0.4726 (0.4782)	0.3451 (0.3437)	0.2357 (0.2296)	0.1482 (0.1403)	0.0838 (0.0768)
	0.4342 (0.4342)	0.3370 (0.3313)	0.2474 (0.2381)	0.1696 (0.1591)	0.1069 (0.0972)	0.0606 (0.0532)
	0.2644 (0.2512)	0.2054 (0.1917)	0.1508 (0.1378)	0.1034 (0.0920)	0.0652 (0.0562)	0.0370 (0.0308)
	0.1233 (0.1110)	0.0956 (0.0847)	0.0701 (0.0609)	0.0480 (0.0407)	0.0303 (0.0249)	0.0172 (0.0136)
The normalizing constant: 3.4571 (3.4460)						
The error distribution:						
0.3475 (0.3492)	0.3418 (0.3439)	0.3198 (0.3220)	0.2772 (0.2791)	0.2031 (0.2029)	0.1261 (0.1239)	0.0704 (0.0679)
The mean squared error: 1.1112 (1.0971)						

### Statistical Linearization

By the method of Booton<sup>14</sup> for the statistical parameters of non-linear systems, the mean squared error for the second-order system is given by

$$\frac{N^2}{2} \left( \frac{\pi}{4} \frac{N^2 T^4}{k^2} + k \right)$$

Table 5. Numerical Solution for a Third-Order System with Symmetry About the Origin

Table 1. Natural Frequencies of a Uniform Beam System with Symmetry About the Origin										
0.031	0.045	0.060	0.078	0.097	$\uparrow \ddot{y}$	0.143	0.077	0.225	0.212	0.171
0.051	0.075	0.102	0.113	0.168	0.207	0.255	0.324	0.320	0.288	0.226
0.078	0.115	0.159	0.212	0.273	0.344	0.435	0.444	0.435	0.396	0.316
0.115	0.172	0.241	0.325	0.422	0.531	0.550	0.553	0.532	0.477	0.378
0.158	0.230	0.313	0.402	0.490	0.486	0.482	0.473	0.448	0.400	0.316
					$\dot{y} = \Delta y$					
					$\uparrow \ddot{y}$					
0.001	0.011	0.031	0.060	0.096	0.137	0.189	0.263	0.275	0.255	0.201
0.026	0.053	0.097	0.158	0.236	0.330	0.454	0.466	0.394	0.293	0.196
0.104	0.181	0.300	0.470	0.701	1.000					
					$\dot{y} = 0$					

and the optimum value of  $k$  as

$$\left(\frac{\pi}{2} N^2 T^4\right)^{\frac{1}{3}}$$

The correct values are respectively

$$\frac{N^2}{2} \left( \frac{N^2 T^4}{k^2} + k \right)$$

and

$$(2 N^2 T^4)^{\frac{1}{3}}$$

For the first-order system, the method of statistical linearization gives

$$\sigma e_{\min}^2 = 0.398; \quad k_{\text{opt}} = 0.6$$

by numerical solution the following is obtained:

$$\sigma e_{\min}^2 = 0.46; \quad k_{\text{opt}} \approx 1.0$$

### Conclusions

The numerical method described in this report is a means by which the analysis of simple non-linear control systems with Gaussian inputs may proceed. Solutions of useful accuracy can be obtained on present sizes of digital computing machines for systems of two or three coordinates, and high accuracy is possible for systems of two coordinates.

Solutions for such simple systems can provide unknown information such as the nature of the distribution of signals in non-linear control systems, and the gain to be expected from non-linear control, the validity of linear approximations, and the accuracy with which a given set of transition probabilities describes a continuous random process.

The writer is indebted to Dr. A.T. Fuller for much helpful supervision.

### References

- <sup>1</sup> EINSTEIN, A. *Ann. Phys., Lpz.* 17 (1905) 549
- <sup>2</sup> KOLMOGOROV, A. N. On analytical methods in probability theory. *Math. Ann.* 104 (1931) 415
- <sup>3</sup> ANDRONOV, A. A., PONTRYAGIN, L. S. and WITT, A. A. On the statistical investigation of dynamical systems. *J. exp. theor. Phys.* 3 (3) (1933) 165
- <sup>4</sup> WANG, M. C. and UHLENBECK, G. E. On the theory of the Brownian motion. *Rev. mod. Phys.* 15 (1944) 165
- <sup>5</sup> CHUANG, K. and KAZDA, L. F. A study of non-linear systems with random inputs. Applications and industry. *Amer. Inst. elect. Engrs* No. 42, May (1959) 100
- <sup>6</sup> PUGACHEV, V. S. *Energ. i Automat.* No. 3 (1961) 46
- <sup>7</sup> BARRETT, J. F. Application of Kolmogorov's equations to randomly distributed automatic control systems. *Proc. 1st IFAC Congr.* 1960
- <sup>8</sup> FLORENTIN, J. J. *Thesis*, University of London, 1960
- <sup>9</sup> FULLER, A. T. The double integrator example quoted by Barrett<sup>7</sup>
- <sup>10</sup> KHAZEN, E. M. Determining the probability distribution for random processes in systems with non-linearities of the piecewise-linear type. *Energ. i Automat.* No. 3 (1961) 58
- <sup>11</sup> SAWARAGI, *et al.* Reports of the Engineering Research Institute, Nos. 68, 79. Kyoto Univ. (1960-61)
- <sup>12</sup> FORSYTHE, G. E. and WASOW, W. R. *Finite Difference Methods for Partial Differential Equations*. 1960. New York and London; Wiley
- <sup>13</sup> CRANK, J. and NICOLSON, P. A practical method for evaluation of solutions of partial differential equations of the heat-conduction type. *Proc. Camb. phil. Soc.* 43 (1947) 50
- <sup>14</sup> BOOTON, R. C. Non-linear control systems with random inputs. *Trans. Inst. Radio Engrs* PGIT 1 (1954)

### DISCUSSION

P. C. PARKS, *Department of Aeronautics, The University, Southampton*

As Dr. Merklinger remarks, there are very few known solutions of the Fokker-Planck-Kolmogorov equation. Those solutions which are known, for example that for linear systems due to Wang and Uhlenbeck (Reference 4 of his paper), the system of Barrett's Figure 3 (Reference 7 in the paper) and the result of Fuller [eqn (36)], take the form

$$\omega = D \exp \left\{ \frac{-V}{N^2} \right\}$$

where  $N^2$  is the power spectral density and  $V$  is a Liapunov function which may be used to prove that the undisturbed system has a stable origin,  $D$  being a normalizing constant.

One new explicit result found by this approach is the phase space steady-state probability distribution for the  $n$ th order linear equation

$$\frac{d^n x}{dt^n} + a_1 \frac{d^{n-1} x}{dt^{n-1}} + a_2 \frac{d^{n-2} x}{dt^{n-2}} + \dots + a_n x = \xi$$

where  $\xi$  is Gaussian white noise. The result<sup>1</sup> is

$$\omega = D \exp \left\{ -\frac{1}{N^2} x' P x \right\}$$

where  $x' = (x, \dot{x}, \ddot{x}, \dots, x^{(n-1)})$ , the phase space state vector,

$$D = \sqrt{\Delta_n \Delta_{n-1}} / (N^2 \pi)^{n/2}$$

and the elements  $P_{ij}$  of  $P$  are defined as follows:

$$P_{ij} = \sum_{k=0}^{n-1} (-1)^{k+n-i} a_k a_{2n-i-j-k+1} \quad \begin{matrix} i+j \text{ even} \\ (i \geq j) \end{matrix}$$

$$P_{ij} = 0 \quad i+j \text{ odd}$$

$$P_{ji} = P_{ij} \quad (i < j)$$

$a_0 = 1$ ,  $a_r = 0$  if  $r > n$ .  $\Delta_n$  and  $\Delta_{n-1}$  are the  $n$ th and  $(n-1)$ th Hurwitz determinants of the  $n$ th order linear equation.

The matrix  $P$  was discovered by Ralston<sup>2</sup> and the quadratic form  $x' P x$  provides also a proof, in two lines, of the Hurwitz stability criterion using Liapunov's second method<sup>1</sup>.

### References

- <sup>1</sup> PARKS, P. C. Automatic Control. *Trans. I.E.E.E.* vol. AC-8, No. 3 (1963) (Correspondence)
- <sup>2</sup> RALSTON, A. Automatic Control. *Trans. Inst. Radio Engrs* vol. AC-7 (July, 1962), pp. 50-51 (Correspondence)

V. S. PUGACHEV, *Inst. of Automatics and Telemechanics of the Ac. Sci. of U.S.S.R., Moscow, I-53, Kalanchevskaya, 15a, U.S.S.R.*

The paper presented by Dr. Merklinger is an interesting extension of studies of non-linear systems with the aid of the theory of Markov random processes. The only remark I should like to make is that it would, perhaps, be preferable to obtain a numerical solution of the integro-differential equation given in my discussion of Dr. J. F. Barrett's paper at the 1st I.F.A.C. Congress, Basle, for the process characteristic function  $g(t; \lambda)$ :

$$\frac{\partial g(\lambda; t)}{\partial t} = \frac{1}{(2\pi)^n} \int_{-\infty}^{\infty} dx \int_{-\infty}^{\infty} e^{i(\lambda - \mu)x} \Phi(\lambda; t_1 x) g(\mu; t) d\mu$$

rather than that of Fokker-Planck-Kolmogorov equation. Since the right member of the above equation is simply the expectation of the random variable  $e^{i\lambda X} \Phi(\lambda; t, X)$  for fixed  $t$ ,  $X = X(t)$  being the Markov

process under study, it is possible to evaluate the integral in the above equation by means of the Monte Carlo method at each stage of the numerical solution. The volume of necessary calculations will then increase less rapidly with the order of the original set of differential equations, than in the case of the numerical solution of the Fokker-Planck-Kolmogorov equation, and this may present some advantages.

K. J. MERKLINGER, *in reply*

Mr. Parks' solution for linear systems is more explicit than any published hitherto.

Professor Pugachev's suggestion to apply numerical methods to the integral equation is quite interesting, and could prove to be more economical. However, since an algorithm for the integral equation would be without physical significance, there might be some difficulty with convergence, boundary conditions, etc.

# Volterra Series Representation of Time-varying Non-linear Systems\*

R. H. FLAKE

## Summary

In this paper the concept of a functional Taylor series is applied directly to the functional relation between the solution and forcing function of a large class of differential equations, yielding the Volterra series. A simple, convenient method is presented for calculating the Volterra kernels in the Volterra series representation of responses with zero and non-zero initial conditions, for a broad class of time-varying non-linear physical systems.

## Sommaire

Dans ce rapport, on utilise le concept d'une série fonctionnelle de Taylor pour représenter la relation fonctionnelle entre la solution et la sollicitation d'une grande classe d'équations différentielles, donnant des séries de Volterra. On montre une méthode simple pour calculer les noyaux de Volterra dans la représentation, par série de Volterra, des réponses avec des conditions initiales nulles et non nulles, pour une large classe de systèmes physiques non-linéaires à coefficients variables avec le temps.

## Zusammenfassung

In diesem Beitrag wird der Begriff der Funktional-(Taylor-)Reihe unmittelbar auf die Funktionalbeziehung zwischen der Lösung und der Anregungsfunktion einer großen Klasse von Differentialgleichungen angewendet; dabei ergeben sich Volterrasche Reihen.

Die Arbeit enthält ein einfaches, bequemes Verfahren zur Berechnung des Volterraschen Kernes in den durch Volterrasche Reihen dargestellten Antworten. Dies gilt für eine große Klasse von zeitveränderlichen nichtlinearen physikalischen Systemen mit sowohl verschwindenden als auch nicht verschwindenden Anfangsbedingungen.

## Introduction

The Volterra series represents, for non-linear systems, the generalization of the concept of a transfer function, which is of primary importance in the analysis and design of linear systems. The Volterra series represents an explicit input-output relation for non-linear systems, and consists of an infinite series composed of terms of the form of convolution integrals. The first term is the convolution integral of the first order kernel and the forcing function of the differential equation describing the system. The first-order kernel is the impulse response of the linear portion of the non-linear differential equation describing the system. The  $n$ th order term is an  $n$ -fold convolution integral containing the  $n$ th order kernel multiplied by an  $n$ th order product of the forcing function.

Historically, this type of series first appeared around 1910 in the studies of Volterra<sup>1</sup> and some of his contemporaries on functional equations. More recently it has been studied by Wiener<sup>2</sup>, Brilliant<sup>3</sup>, George<sup>4</sup>, McFee<sup>5</sup>, Blackman<sup>6</sup>, and others. Zadeh<sup>7, 8</sup> considers a generalization of the Volterra series.

\* This work contains part of the results of a dissertation submitted by R. H. Flake in partial fulfilment of the requirements for the degree of Doctor of Science in Engineering at Washington University.

Volterra, Brilliant, and Blackman present proofs showing that, under certain conditions, a functional  $y(t) = T[x(t)]$  can be approximated to any desired degree of accuracy by a finite series of the form

$$y(t) = \int_0^t h_1(\tau) x(t-\tau) d\tau + \int_0^t \int_0^t h_2(\tau_1, \tau_2) x(t-\tau_1) x(t-\tau_2) d\tau_1 d\tau_2 + \dots + \int_0^t \dots \int_0^t h_n(\tau_1, \dots, \tau_n) x(t-\tau_1) x(t-\tau_2) \dots x(t-\tau_n) d\tau_1 \dots d\tau_n + \dots \quad (1)$$

Such a functional is called continuous<sup>1, 3, 6</sup>. It is easy to show that the functional relation between the solution and the forcing function of a non-linear differential equation with constant coefficients, satisfying the Lipschitz condition, is continuous. If  $T[x(t)]$  can be represented exactly by a converging infinite series of the form of eqn (1), it is called analytic<sup>1</sup>. Volterra shows that eqn (1) can then be interpreted as a functional generalization of the Taylor series expansion for the analytic functional. Conditions for convergence of the Volterra series are considered by Volterra and Brilliant.

One of the main problems in the application of the theory is the explicit determination of the kernels appearing in the Volterra series. Both analytical and experimental procedures for determining these kernels for time-invariant non-linear systems have been investigated<sup>4, 5, 9</sup> and the last two sections of this paper show that the problem of calculating the Volterra kernels associated with time-varying non-linear systems may be approached in basically the same manner as the method for time-invariant systems described by Flake<sup>9</sup>. As in the previous paper<sup>9</sup>, a Volterra series representing the response of a non-linear system with zero initial conditions is first considered here, and then modified to represent a response with arbitrary initial conditions. In the first part of this paper, Volterra's treatment<sup>1</sup> of the functional Taylor series for an abstract analytic functional is applied directly to the functional relation between the solution of a forced non-linear differential equation and the forcing function.

## The Functional Taylor Series

Define the functional,  $\phi = T[y]$ , by the differential equation<sup>†</sup>,

$$\phi(t) = \frac{dy}{dt} - f(t, y) \quad (2)$$

† The results in the following are easily generalized to a system of differential equations which can be written in the form of eqn (2) by using matrix notation.

where  $y(t)$  is the set of continuous functions with continuous first derivatives defined over the interval  $0 \leq t \leq a$ , with  $y(0) = 0$ .  $f(t, y)$  is a continuous function of  $t$  and an entire function of  $y$  for each value of  $t$ , with  $f(t, 0) = 0$ . The solution,  $y(t)$ , of eqn (2) is given formally by  $T^{-1}[\phi(t)] = F[\phi(t)]$ .

Now suppose the forcing function,  $\phi(t)$ , is changed by an amount  $\varepsilon x(t)$ , where  $x(t)$  is a continuous function and  $\varepsilon$  is a parameter. Then the solution becomes

$$y_\varepsilon(t) = F[\phi(t) + \varepsilon x(t)]$$

If  $t$ ,  $\phi(t)$ , and  $x(t)$  are assumed fixed, and  $t$  is sufficiently small, then  $y_\varepsilon$  may be considered an ordinary function of  $\varepsilon$ . Assuming  $y_\varepsilon$  has continuous derivatives up to  $n$ th order with respect to  $\varepsilon$  on the interval  $0 \leq \varepsilon \leq b$ , then, by Taylor's theorem (actually,  $y_\varepsilon(t)$  is an analytic function of  $\varepsilon$  for  $t$  sufficiently small<sup>10</sup>),

$$y_\varepsilon = y_\varepsilon|_{\varepsilon=0} + \frac{\partial y_\varepsilon}{\partial \varepsilon} \Big|_{\varepsilon=0} \varepsilon + \frac{\partial^2 y_\varepsilon}{\partial \varepsilon^2} \Big|_{\varepsilon=0} \frac{\varepsilon^2}{2!} + \dots + \frac{\partial^{n-1} y_\varepsilon}{\partial \varepsilon^{n-1}} \Big|_{\varepsilon=0} \frac{\varepsilon^{n-1}}{(n-1)!} + \frac{\partial^n y_\varepsilon}{\partial \varepsilon^n} \Big|_{\varepsilon=0} \frac{\varepsilon^n}{n!} \quad (3)$$

where  $0 \leq \theta \leq \varepsilon \leq b$ . The terms in eqn (3) may be interpreted as follows:  $y_\varepsilon(t)$  is the solution of the differential equation

$$\frac{dy_\varepsilon}{dt} = f(t, y_\varepsilon(t)) + \phi(t) + \varepsilon x(t) \quad (4)$$

satisfying the initial conditions  $y_\varepsilon(0) = 0$ . The first term on the right side of eqn (3) is the solution of eqn (4) satisfying zero initial conditions when  $\varepsilon = 0$ , and is therefore a solution of eqn (2). The second term on the right side of eqn (3) in  $\varepsilon$  multiplied by the quantity

$$\frac{\partial y_\varepsilon}{\partial \varepsilon} \Big|_{\varepsilon=0} = \lim_{\varepsilon \rightarrow 0} \frac{F[\phi(t) + \varepsilon x(t)] - F[\phi(t)]}{\varepsilon} \quad (5)$$

which is called the functional differential of  $F$  with respect to  $x(t)$ , evaluated at  $\phi(t)$  and denoted by  $\delta F[\phi, x]$ . This quantity will presently be explicitly calculated.

The higher order functional differentials appearing in eqn (3) are obvious generalizations of (5). Thus, eqn (3) is a functional generalization of the Taylor series for a solution of eqn (4), expanded about a solution of eqn (2). Note that, if a solution of eqn (2) satisfying certain initial conditions is known, then evaluation of eqn (3) would yield the solution of eqn (4) satisfying the same initial conditions. In particular, if  $\phi(t) = 0$ , then the solution of eqn (2) satisfying zero initial conditions is  $y(t) = 0$ , and therefore an evaluation of eqn (3) would yield the solution of

$$\frac{dy}{dt} = f(t, y) + \varepsilon x(t)$$

satisfying the initial condition,  $y(0) = 0$ .

### Calculation of the Functional Differential

The functional differentials appearing in eqn (3) must of course be calculated in order to apply the functional Taylor series. Consider the first differential given by eqn (5).  $F[\phi(t) + \varepsilon x(t)]$  is a solution of eqn (4), and  $F[\phi(t)]$  is a solution of

eqn (2), both satisfying zero initial conditions. Let  $l y = F[\phi(t) + \varepsilon x(t)] - F[\phi(t)]$ . Then<sup>11</sup>

$$\frac{d}{dt}(\Delta y) = f(t, y + \Delta y) - f(t, y) + \varepsilon x(t)$$

or

$$\frac{d}{dt}(\Delta y) = \Delta y \int_0^1 f_y(t, y + \eta \Delta y) d\eta + \varepsilon x(t)$$

where  $f_y$  is the derivative of  $f(t, y)$  with respect to  $y$ .

Now let  $A(t) = \int_0^1 f_y(t, y + \eta \Delta y) d\eta$ , and consider the solution  $l(t, t')$  of

$$\frac{dl}{dt} = -A(t)l$$

where  $l(t', t') = 1$ . Then, multiplying both sides of the differential equation above for  $\Delta y$  by  $l$  yields

$$l \frac{d}{dt} \Delta y = -\Delta y \frac{dl}{dt} + \varepsilon l x$$

or

$$l(t, t') \Delta y(t) \Big|_0^{t'} = \varepsilon \int_0^{t'} l(\tau, t') x(\tau) d\tau$$

Finally,  $\Delta y(t') = \varepsilon \int_0^{t'} l(\tau, t') x(\tau) d\tau$ , since  $\Delta y(0) = 0$ , and  $l(t', t') = 1$ .

Now

$$\begin{aligned} \delta F[\phi, x] &= \lim_{\varepsilon \rightarrow 0} \frac{\Delta y(t)}{\varepsilon} \\ &= \lim_{\varepsilon \rightarrow 0} \int_0^t l(\tau, t) x(\tau) d\tau \end{aligned}$$

Let  $l_0$  be defined as the solution of

$$\frac{dl_0}{dt} = -A_0(t)l_0 \quad (6)$$

with  $l_0(t', t') = 1$ , where  $A_0(t) = \lim_{\varepsilon \rightarrow 0} \int_0^1 f_y(t, y + \eta \Delta y) d\eta$ . But

$\lim_{\varepsilon \rightarrow 0} \Delta y = 0$ . Thus,  $A_0(t) = \int_0^1 d\eta f_y(t, y)$ , or  $A_0(t) = f_y(t, y(t))$ ,  $y(t)$  being the solution of eqn (2) satisfying zero initial conditions. Finally,

$$\delta F[\phi, x] = \int_0^t l_0(\tau, t) x(\tau) d\tau \quad (7)$$

Before evaluating eqn (7) for a particular example, the second differential will be calculated.

Consider

$$\delta^2 F[\phi, x] = \lim_{\varepsilon \rightarrow 0} \frac{\delta F[\phi + \varepsilon x, x] - \delta F[\phi, x]}{\varepsilon}$$

Now  $\delta F[\phi, x]$  is given by eqn (7); and similarly,

$$\delta F[\phi + \varepsilon x, x] = \int_0^t l_1(\tau, t) x(\tau) d\tau$$

where  $l_1(t, t')$  is the solution of

$$\frac{dl_1}{dt} = -f_y(t, y_\varepsilon) l_1$$

satisfying the initial conditions  $l_1(t', t') = 1$ . Thus

$$\delta^2 F[\phi, x] = \int_0^t \lim_{\varepsilon \rightarrow 0} \left( \frac{l_1(\tau, t) - l_0(\tau, t)}{\varepsilon} \right) x(\tau) d\tau \quad (8)$$

Let  $\Delta l = l_1 - l_0$ . Then

$$\frac{d}{dt}(\Delta l) = -f_y(t, y_\varepsilon) l_1 + f_y(t, y) l_0$$

or

$$\frac{d}{dt}(\Delta l) = -f_y(t, y_\varepsilon) \Delta l - [f_y(t, y + \Delta y) - f_y(t, y)] l_0 \quad (9)$$

Now,  $f_y(t, y + \Delta y) - f_y(t, y) = \Delta y \int_0^1 f_{yy}(t, y + \eta \Delta y) d\eta$ .

Then eqn (9) becomes

$$\frac{d}{dt}(\Delta l) = -B(t) \Delta l - C(t) l_0 \quad (10)$$

where  $B(t) = f_y(t, y_\varepsilon(t))$ , and  $C(t) = \Delta y \int_0^1 f_{yy}(t, y + \eta \Delta y) d\eta$ .

Consider the solution of

$$\frac{dm}{dt} = B(t) m, \text{ with } m(t, t'')|_{t=t''} = 1$$

Then, multiplying eqn (10) by  $m(t, t'')$  yields

$$m \frac{d}{dt}(\Delta l) = -\Delta l \frac{dm}{dt} - C(t) m(t, t'') l_0(t, t')$$

Integrating yields

$$m(t, t'') \Delta l(t, t') \Big|_{t''}^{t'} = - \int_{t''}^{t'} m(\tau, t'') C(\tau) l_0(\tau, t') d\tau$$

or

$$\Delta l(t'', t') = \int_{t''}^{t'} m(\tau, t'') C(\tau) l_0(\tau, t') d\tau$$

Thus,

$$\lim_{\varepsilon \rightarrow 0} \frac{\Delta l(t'', t)}{\varepsilon} = \int_{t''}^{t'} d\tau m_0(\tau, t) l_0(\tau, t') f_{yy}(\tau, y(\tau)) \delta F[\phi(\tau)]$$

and, from eqn (8)

$$\boxed{\text{Eqn (11)}}^*$$

Finally, using eqn (7)

$$\boxed{\text{Eqn (11a)}}^\dagger$$

where  $l_0(t, t')$  is defined by eqn (6), and  $m_0(t, t'')$  is the solution of

$$\frac{dm_0}{dt} = f_y(t, y(t)) m_0, m_0(t, t'')|_{t=t''} = 1 \quad (12)$$

The higher order differentials of  $F$  may be calculated in a similar manner.

Eqn (11a) may be written in the form

$$\delta^2 F[\phi, x] = 2! \int_0^t \int_0^t h_2(t, \tau, \eta) x(\tau) x(\eta) d\tau d\eta \quad (13)$$

if  $h_2(t, \tau, \eta)$  is defined ( $u(t)$  is the unit step function, which is equal to one for  $t > 0$ , and zero for  $t < 0$ ) as

$$\boxed{\text{Eqn (13a)}}^\neq$$

or

$$\boxed{\text{Eqns (14, 14a)}}^\S$$

Volterra<sup>1</sup> writes the  $n$ th order differential in the form

$$\begin{aligned} \delta^n F[\phi, x] \\ = n! \int_0^t \dots \int_0^t h_n(t, \tau_1, \dots, \tau_n) x(\tau_1) \dots x(\tau_n) d\tau_1 \dots d\tau_n \end{aligned}$$

Thus, eqn (3) may be written as

$$\begin{aligned} y_\varepsilon(t) = & \varepsilon \int_0^t h_1(t, \tau) x(\tau) d\tau \\ & + \varepsilon^2 \int_0^t \int_0^t h_2(t, \tau_1, \tau_2) x(\tau_1) x(\tau_2) d\tau_1 d\tau_2 + \dots \\ & + \frac{\varepsilon^n}{n!} \delta^n F[\phi(t) + \theta x(t)] \end{aligned} \quad (15)$$

\* Eqn (11)

$$\delta^2 F[\phi, x] = \int_0^t d\tau x(\tau) \int_\tau^t d\theta m_0(\theta, \tau) l_0(\theta, t) f_{yy}(\theta, y(\theta)) \delta F[\phi(\theta)] \quad (11)$$

† Eqn (11a)

$$\delta^2 F[\phi, x] = \int_0^t d\tau x(\tau) \int_\tau^t d\theta m_0(\theta, \tau) l_0(\theta, t) f_{yy}(\theta, y(\theta)) \int_0^\theta d\eta l_0(\eta, \theta) x(\eta) \quad (11a)$$

≠ Eqn (13a)

$$h_2(t, \tau, \eta) = \frac{1}{2!} \int_\tau^t m_0(\theta, \tau) l_0(\theta, t) f_{yy}(\theta, y(\theta)) u(\theta - \eta) l_0(\eta, \theta) d\theta \quad (13a)$$

§ Eqns (14, 14a)

$$\text{or } h_2(t, \tau, \eta) = \begin{cases} \frac{1}{2!} \int_\tau^t m_0(\theta, \tau) l_0(\theta, t) f_{yy}(\theta, y(\theta)) l_0(\eta, \theta) d\theta & \text{if } \tau > \eta \\ \frac{1}{2!} \int_\eta^\tau m_0(\theta, \tau) l_0(\theta, t) f_{yy}(\theta, y(\theta)) l_0(\eta, \theta) d\theta & \text{if } \tau < \eta \end{cases} \quad (14)$$

$$(14a)$$

where  $0 \leq \theta \leq \varepsilon$ . If the number of terms is allowed to become arbitrarily large, then the infinite series is called the Volterra series. Note that the last term in eqn (15) may be used to estimate the error in the approximate evaluation of  $y_\varepsilon$  for a series composed of a finite number of terms. An upper boundary on the remainder term may be calculated, provided an upper boundary on  $y_\varepsilon(t)$  is known.

*Example*—Consider the differential equation

$$\frac{dy_\varepsilon}{dt} + y_\varepsilon + Ky_\varepsilon^2 = \varepsilon x(t) \quad (16)$$

The first two terms in the functional expansion of  $y_\varepsilon$  about zero forcing function ( $\phi(t) = 0$ ), satisfying zero initial conditions, are now calculated.  $l_0(t, t')$  is the solution of eqn (6), which in this example becomes

$$\frac{dl_0}{dt} = l_0, \text{ with } l_0(t', t') = 1$$

since  $A_0(t) = -1$ . Thus,  $l_0(t, t') = e^{t-t'}$  for  $t \leq t'$ , and eqn (7) becomes

$$\delta F[0, x] = \int_0^t e^{\tau-t} x(\tau) d\tau \quad (17)$$

or  $h_1(t, \tau) = e^{\tau-t}$ .  $m_0(t, t'')$  is the solution of eqn (12), thus

$$\frac{dm_0}{dt} = -m_0, \text{ with } m_0(t'', t'') = 1$$

Therefore,  $m_0 = e^{-(t-t'')}$ , and eqn (14) becomes

$$\begin{aligned} h_2(t, \tau_1, \tau_2) &= \frac{1}{2!} \int_{\tau_1}^t m_0(\theta, \tau_1) l_0(\theta, t) f_{yy}(\theta, y(\theta)) l_0(\tau_2, \theta) d\theta \\ &= \frac{1}{2!} (-2K) \int_{\tau_1}^t e^{-(\theta-\tau_1)} e^{(\theta-t)} e^{(\tau_2-\theta)} d\theta \\ h_2(t, \tau_1, \tau_2) &= K [e^{-(t-\tau_1)} e^{-(t-\tau_2)} - e^{-(t-\tau_2)}] \text{ for } \tau_1 > \tau_2 \end{aligned} \quad (18)$$

Similarly, eqn (14a) yields

$$h_2(t, \tau_1, \tau_2) = K [e^{-(t-\tau_1)} e^{-(t-\tau_2)} - e^{-(t-\tau_1)}] \text{ for } \tau_1 < \tau_2 \quad (18a)$$

Thus, if  $x(t)$  is a unit step function, then the first two terms in eqn (3) become

$$y_\varepsilon(t) = \varepsilon \int_0^t e^{\tau-t} d\tau + \varepsilon^2 K \int_0^t \int_0^t e^{-(t-\tau_1)} e^{-(t-\tau_2)} d\tau_1 d\tau_2$$

$$+ \varepsilon^2 K \int_0^t d\tau_1 \left[ \int_0^{\tau_1} d\tau_2 e^{-(t-\tau_2)} + \int_{\tau_1}^t d\tau_2 e^{-(t-\tau_1)} \right] + \dots$$

or

$$y_\varepsilon(t) = \varepsilon(1 - e^{-t}) - K\varepsilon^2(1 - 2te^{-t} - e^{-2t}) + \dots \quad (19)$$

$h_1(t, \tau)$ ,  $h_2(t, \tau_1, \tau_2)$  and the higher order kernels in the Volterra series are called Volterra kernels. A more direct and convenient method for calculating the Volterra kernels is given in the next section.

### Determining the Volterra Kernels

A simple, convenient technique is introduced in this section for calculating the Volterra kernels associated with physical systems which can be described by ordinary differential equations of the form

$$L[y(t)] + F[t, y(t), y'(t), \dots, y^{(n-1)}(t)] = x(t) \quad (20)$$

where  $L$  is a linear differential operator with variable coefficients, and  $F$  is a polynomial in  $y(t)$ ,  $y'(t)$ ,  $\dots$ ,  $y^{(n-1)}(t)$ , beginning with terms of order no lower than the second. (The treatment could be applied to the more general vector form of eqn (2). However, equations of the form of eqn (20) are considered here for the sake of simplifying the notation.) The coefficients of the polynomial  $F$  are, in general, continuous functions of  $t$ .  $n$  is the order of the highest derivative of  $y(t)$  appearing in  $L[y(t)]$ .  $x(t)$  is a continuous function.

The basis of the technique is the assumption that the particular solution of this class of forced non-linear differential equations can be represented by a converging infinite series of the form

$$\boxed{\text{Eqn (20a)}}^*$$

where the kernels  $h_n(t, \tau_1, \dots, \tau_n)$  are continuous functions, symmetric in  $\tau_i$ . The particular solution for an  $n$ th order differential equation represented by eqn (20a) is, of course, the one for which the solution and its first  $n-1$  derivatives are all zero at  $t=0$ . A modified form of eqn (20a) is later used to represent solution with non-zero initial conditions.

The proposed procedure for calculating the kernels in a Volterra series solution of eqn (20) is to substitute eqn (20a) into eqn (20) and equate the coefficients of the different functionals of  $x(t)$ . Since  $x(t)$  is an arbitrary forcing function, then the coefficients of each functional of  $x(t)$  must vanish if eqn (20a) is to be a solution of the differential equation. The kernels in the Volterra series solution are uniquely determined by the resulting set of equations. Detailed calculation of the kernels for a simple example will clarify the procedure.

Let the system of interest be described by the first order differential equation

$$A(t) \frac{dy}{dt} + B(t)y + Ky^2 = x(t) \quad (21)$$

\* Eqn (20a):

$$y(t) = \int_0^t h_1(t, \tau) x(\tau) d\tau + \iint_0^t h_2(t, \tau_1, \tau_2) x(\tau_1) x(\tau_2) d\tau_1 d\tau_2 + \dots + \int_0^t \dots \int_0^t h_n(t, \tau_1, \dots, \tau_n) x(\tau_1) \dots x(\tau_n) d\tau_1 \dots d\tau_n + \dots \quad (20a)$$



Substitution of eqn (20a) into eqn (21) and equating the coefficients of the different functionals of  $x(t)$  yields

$$\boxed{\text{Eqn (22-27)}}^*$$

These equations may be solved sequentially for the kernels. Eqns (22) and (23) yield

$$h_1(t, \tau) = \frac{1}{A(\tau)} e^{-\int_{\tau}^t \frac{B(\theta)}{A(\theta)} d\theta} \quad \text{for } t > \tau$$

$$= 0 \quad \text{for } t < \tau \quad (28)$$

$h_2(t, \tau_1, \tau_2)$  is determined from eqn (25) and the boundary conditions determined from eqn (24). Note that eqn (25) is a non-homogeneous form of eqn (23). Thus  $h_1(t, \tau)$  can be used as a one-sided Green's function<sup>12, 13</sup> for eqn (25). Therefore

$$h_2(t, \tau_1, \tau_2) = -K \int_0^t d\theta G(t, \theta) h_1(\theta, \tau_1) h_1(\theta, \tau_2) \quad (29)$$

where  $G(t, \tau) = h_1(t, \tau)$  for  $t > \tau$ . Eqn (29) is a solution of eqn (25) in terms of the previously determined first order kernel, and it satisfies the boundary conditions given by eqn (24). The  $n$ th order kernel  $h_n(t, \tau_1, \dots, \tau_n)$  is determined in terms of the kernels of order less than  $n$  from eqns (26) and (27). That is

$$h_n(t, \tau_1, \dots, \tau_n) = -K \int_0^t d\theta G(t, \theta) \left[ \sum_{j=1}^{n-1} h_j(\theta, \tau_1, \dots, \tau_j) h_{n-j}(\theta, \tau_{j+1}, \dots, \tau_n) \right]$$

$$\text{for } n=2, 3, \dots \quad (30)$$

Thus the higher order kernels may be obtained from repeated use of eqn (30).

In general, the Volterra kernels associated with a differential equation of the form of eqn (20) are determined by the equations resulting from equating the coefficients of the different functionals of  $x(t)$  that appear when eqn (20a) is substituted into the differential equation. The first order kernel is recognized to be the impulse response of the linear portion of the differential equation. The higher order kernels are solutions of non-homogeneous linear partial differential equations, satisfying associated boundary conditions. The solution of the equation determining the  $n$ th order kernel may be expressed in integral form as the product of the first order kernel, used as a one-sided Green's function, and kernels of order less than  $n$ . Thus, if the first order kernel is known, then the determination of the analytical form of the higher order kernels depends upon being able successively to evaluate integrals containing previously determined lower order kernels. (If the linear portion of the differential equation has constant coefficients, then the first order kernel may be obtained by routine calculation.) Note, however, that a closed form, or, indeed, any exact solution for the first order kernel, may not be obtainable, since in general it is the impulse response of a linear differential equation with variable coefficients. Nevertheless, it is often possible to obtain a suitable approximate expression for the first order kernel, which could then be used to calculate approximate higher order kernels. The resulting approximate Volterra series solution may be sufficiently accurate for the purposes of the analysis.

#### Non-zero Initial Conditions

The general solution of a differential equation of the form of eqn (20) may be represented by a Volterra series modified by the introduction of a pseudo-forcing function.

Consider first the generalization of eqn (20a) to represent a solution of eqn (20) which has some arbitrary non-zero value

\* Eqns (22-27):

Functional	Coefficient	
$x(t)$	$A(t) h_1(t, t) = 1$	(22)

$\int_0^t d\tau x(\tau)$	$A(t) \frac{\partial}{\partial t} h_1(t, \tau) + B(t) h_1(t, \tau) = 0$	(23)
--------------------------	---	------

$x(t) \int_0^t d\tau x(\tau)$	$A(t) [h_2(t, \tau, t) + h_2(t, t, \tau)] = 0$	(24)
-------------------------------	--	------

$\int_0^t \int_0^t d\tau_1 d\tau_2 x(\tau_1) x(\tau_2)$	$A(t) \frac{\partial}{\partial t} h_2(t, \tau_1, \tau_2) + B(t) h_2(t, \tau_1, \tau_2) + K h_1(t, \tau_1) h_1(t, \tau_2) = 0$	(25)
---	---	------

$x(t) \int_0^t \dots \int_0^t \prod_{i=1}^n d\tau_i x(\tau_i)$	$A(t) \sum_{j=1}^n h_n(t, \tau_1, \dots, \tau_{j-1}, t, \tau_{j+1}, \dots, \tau_n) = 0$	(26)
--	---	------

$\int_0^t \dots \int_0^t \prod_{i=1}^n d\tau_i x(\tau_i)$	$A(t) \frac{\partial}{\partial t} h_n(t, \tau_1, \dots, \tau_n) + B(t) h_n(t, \tau_1, \dots, \tau_n) + K \sum_{j=1}^{n-1} h_j(\tau_1, \dots, \tau_j) h_{n-j}(\tau_{j+1}, \dots, \tau_n) = 0$	(27)
---	--	------

initially, and whose first  $n - 1$  derivatives are initially zero. This may easily be accomplished by introducing an auxiliary variable which is the difference between the desired solution and its initial value. This new variable will satisfy a differential equation of the form of eqn (20), if the terms in the new differential equation which do not depend on  $y(t)$  are included in the forcing function. The new variable may then be represented by a series of the form of eqn (20a). Finally, the desired solution is the sum of the series for the auxiliary variable and an arbitrary constant. The generalization of the Volterra series to represent a solution for a second or higher order differential equation satisfying non-zero initial conditions on its derivatives, is perhaps not as obvious. A particular example will first be presented, and then the procedure for calculating the modified Volterra series representing the general solution of an  $n$ th order differential equation of the form of eqn (20) will be stated.

Consider the second order non-linear time-varying differential equation

$$\frac{d^2 y}{dt^2} + t \frac{dy}{dt} + y + y^2 = x(t) \quad (31)$$

A procedure for calculating a modified Volterra series form of solution satisfying the initial conditions

$$\begin{aligned} y(0) &= 0 \\ \left. \frac{dy}{dt} \right|_{t=0} &= y_1 \end{aligned} \quad (32)$$

is now shown.

Define  $f(t)$  by the relation

$$\frac{df}{dt} + f = x(t) \quad (33)$$

and assume a solution for eqn (31) of the form

$$\begin{aligned} y(t) &= \int_0^t h_1(t, \tau) f(\tau) d\tau \\ &+ \int_0^t \int_0^t h_2(t, \tau_1, \tau_2) f(\tau_1) f(\tau_2) d\tau_1 d\tau_2 + \dots \end{aligned} \quad (34)$$

Then substitution of eqns (33) and (34) into eqn (31) and equating the different functionals of  $f(t)$  as before yields

$$h_1(t, t) = 1 \quad (35)$$

$$\frac{\partial}{\partial t} h_1(t, \tau) \big|_{\tau=t} = 1 - t \quad (36)$$

$$\frac{\partial^2 h_1(t, \tau)}{\partial t^2} + t \frac{\partial}{\partial t} h_1(t, \tau) + h_1(t, \tau) = 0 \quad (37)$$

$$h_2(t, \tau, t) + h_2(t, t, \tau) = 0 \quad (38)$$

$$\frac{\partial}{\partial t} h_2(t, \tau_1, \tau) \big|_{\tau_1=t} + \frac{\partial}{\partial t} h_2(t, \tau, \tau_1) \big|_{\tau_1=t} = 0 \quad (39)$$

$$\frac{\partial^2}{\partial t^2} h_2(t, \tau_1, \tau_2) + t \frac{\partial}{\partial t} h_2(t, \tau_1, \tau_2)$$

$$+ h_2(t, \tau_1, \tau_2) + h_1(t, \tau_1) h_1(t, \tau_2) = 0$$

$$\vdots \quad \quad \quad \vdots$$

$$(40) \quad \text{The author deeply appreciates the numerous helpful discussions with Professor John Zaborszky during the course of this study.}$$

The solution of eqn (37) satisfying the initial conditions given by eqns (35) and (36) is

$$h_1(t, \tau) = e^{-t^2/2} \int_{\tau}^t e^{r^2/2} dr + e^{-t^2/2} e^{\tau^2/2} \quad (41)$$

The second and higher order kernels may be expressed in terms of a one-sided Green's function and lower order kernels. The one-sided Green's function is determined from the impulse response of the linear portion of eqn (32). From eqn (38), (39), and (40) it follows that

$$h_2(t, \tau_1, \tau_2) = - \int_0^t d\theta G(t, \theta) h_1(\theta, \tau_1) h_1(\theta, \tau_2) \quad (42)$$

where  $G(t, \tau)$  is given by

$$G(t, \tau) = e^{-t^2/2} \int_{\tau}^t e^{r^2/2} dr, \quad \text{for } t > \tau \quad (43)$$

and  $h_1(t, \tau)$  is the previously determined first order kernel. The higher order kernels may be determined in the same manner. It is easily seen by inspection that the modified Volterra series for  $y(t)$  will satisfy eqn (31) and the desired initial conditions, if  $f(0) = y_1$ .

The proposed procedure for obtaining a modified Volterra series representing the general solution of eqn (20) is first to introduce an auxiliary variable,  $u(t)$ , which has zero initial value and differs from the desired solution by a constant.  $u(t)$  will then satisfy a differential equation of the form of eqn (20), if all terms which are independent of  $u(t)$  are included in the forcing function. A solution of the auxiliary differential equation can be expressed as

$$\begin{aligned} u(t) &= \int_0^t h_1(t, \tau) f(\tau) d\tau \\ &+ \int_0^t \int_0^t h_2(t, \tau_1, \tau_2) f(\tau_1) f(\tau_2) d\tau_1 d\tau_2 + \dots \end{aligned} \quad (44)$$

where  $f(t)$  is the general solution of

$$\frac{d^{(n-1)}}{dt^{(n-1)}} f(t) + \dots + \frac{df}{dt}(t) + f(t) = z(t) \quad (45)$$

and  $z(t)$  is the forcing function of the auxiliary differential equation. If the impulse response of the linear portion of the auxiliary differential equation can be found, then it can be used as a one-sided Green's function to calculate the higher order kernels in eqn (44), by the procedure previously described. Finally, the general solution of eqn (20) will be

$$\begin{aligned} y(t) &= C + \int_0^t h_1(t, \tau) f(\tau) d\tau \\ &+ \int_0^t \int_0^t h_2(t, \tau_1, \tau_2) f(\tau_1) f(\tau_2) d\tau_1 d\tau_2 + \dots \end{aligned}$$

where  $C$  is an arbitrary constant.

## References

- <sup>1</sup> VOLTERRA, V. *Theory of Functionals and of Integral and Integro-Differential Equations*. 1930. London; Blackie
- <sup>2</sup> WIENER, N. *Nonlinear Problems in Random Theory*. 1958. New York; Technology Press and Wiley
- <sup>3</sup> BRILLIANT, M. B. Theory of the analysis of nonlinear systems. *Technical Report 345*, Research Laboratory of Electronics, Massachusetts Institute of Technology (1958)
- <sup>4</sup> GEORGE, D. A. Continuous nonlinear systems. *Technical Report 355*, Research Laboratory of Electronics, Massachusetts Institute of Technology (1959)
- <sup>5</sup> MCFEE, R. Determining the response of nonlinear systems to arbitrary inputs. *Trans. Amer. Inst. elect. Engrs*, Paper No. 61-114 (1960)
- <sup>6</sup> BLACKMAN, J. The representation of nonlinear networks. *Syracuse University Research Institute, Report No. 81560*, for Air Force Cambridge Research Center
- <sup>7</sup> ZADEH, L. A. Nonlinear multipoles. *Proc. Nat. Acad. Sci. Wash.* 39 (1953), 274
- <sup>8</sup> ZADEH, L. A. A contribution to the theory of nonlinear systems. *J. Franklin Inst.* 255 (1953)
- <sup>9</sup> FLAKE, R. H. Volterra series representation of nonlinear systems. *Trans. Amer. Inst. elect. Engrs*, Paper No. 62-1189 (1962)
- <sup>10</sup> CODDINGTON, and LEVINSON *Theory of Ordinary Differential Equations*. 1955. New York; McGraw-Hill
- <sup>11</sup> BLISS, G. A. Differential equations containing arbitrary functions. *Trans. Amer. math. Soc.* 21 (1920), 79
- <sup>12</sup> MILLER, K. S. Properties of impulsive response and Green's functions. *Trans. Inst. Radio Engrs* CT-2 (1955)
- <sup>13</sup> MILLER, K. S. The one-sided Green's function. *J. Appl. Phys.* 22 (1951)

## DISCUSSION

J. F. BARRETT, *Mechanical Engineering Department, Birmingham University, Birmingham, England*

I would like to point out an alternative method of calculating the Volterra series expansion for certain non-linear differential equations. This method is applicable to the examples considered in Dr. Flake's paper. The method is based on the conversion of the differential equation into a non-linear integral equation followed by series solution. When applicable it gives the Volterra series with little calculation. It has the advantage that it permits proof of the Volterra series, which is a very important point since convergence is directly related to stability. To illustrate the method consider the first example of Dr. Flake's paper

$$\dot{y} + y + Ky^2 = \varepsilon x(t), t \geq 0$$

Then

$$\dot{y} + y = \varepsilon x(t) - Ky^2$$

(left-hand side linear)

$$y(t) = y_0 e^{-t} + \varepsilon \int_0^t e^{-(t-\tau)} x(\tau) d\tau - K \int_0^t e^{-(t-\tau)} y^2(\tau) d\tau$$

since the solution of

$$\dot{y} + y = z(t), t \geq 0$$

is

$$y(t) = y_0 e^{-t} + \int_0^t e^{-(t-\tau)} z(\tau) d\tau$$

For simplicity take  $y_0 = 0$

$$y(t) = \varepsilon \int_0^t e^{-(t-\tau)} x(\tau) d\tau - K \int_0^t e^{-(t-\tau)} y^2(\tau) d\tau \quad (1)$$

This is a non-linear integral equation which can be solved in power series, e.g., by successive approximation.

First approximation:

$$y(t) = \varepsilon \int_0^t e^{-(t-\tau)} x(\tau) d\tau + \dots$$

Second approximation:

$$y(t) = \varepsilon \int_0^t e^{-(t-\tau)} x(\tau) d\tau - K \varepsilon^2 \int_0^t e^{-(t-\tau)} \left( \int_0^{\tau} e^{-(\tau-\tau_1)} x(\tau_1) d\tau_1 \right)^2 d\tau + \dots \quad (2)$$

This is the Volterra series.

## Proof of Convergence

Consider the algebraic equation

$$Y = |\varepsilon| X + |K| Y^2 \quad (1a)$$

similar in form to (1).

Similar series solution of this equation would give

$$Y = |\varepsilon| X + |K| |\varepsilon|^2 X^2 + \dots \quad (2a)$$

This is just the series expansion of the solution

$$Y = \frac{1 - \sqrt{1 - 4|\varepsilon|X|K|}}{2|\varepsilon||K|}$$

of the equation (1a) and so is convergent if  $4|\varepsilon||X||K| < 1$  i.e.  $|X| < 1/4|\varepsilon||K|$ .

Now it can easily be shown that every term of series (1) is in absolute magnitude less than or equal to the corresponding term in series (1a) if  $\max_{t_1 > 0} |X(t_1)| < X$  in which case also

$$\max_{t_1 > 0} |y(t_1)| < Y$$

Hence if

$$\|x\| X = \max_{t_1 > 0} |x(t_1)| < \frac{1}{4|\varepsilon||K|},$$

then the Volterra series (1a) is convergent and also

$$\max_{t_1 > 0} |y(t_1)| < \frac{1 - \sqrt{1 - 4|\varepsilon|\|x\||K|}}{2|\varepsilon||K|}.$$

This is also a statement of stability of the system.

## R. H. FLAKE, in reply

I want to thank Dr. Barrett for suggesting his interesting alternate method for calculating the Volterra series. The related investigation of the convergence of the series is certainly important. However, I am not sure that I agree with the statement that the suggested method, when applicable, will give the Volterra series with little calculation, since this type of computation will generally become very laborious after only a few iterations. In Reference 9, the kernels in the Volterra series solution for this example are calculated using the simple procedure suggested in this paper. I believe an examination of the two techniques applied to the same example will indicate that the procedure discussed in the second half of the paper is computationally much less involved than the suggested iterative procedure.

P. ALPER, *Elec. Lab., T.H. Kanaalweg Z B, Delft, Holland*

In this excellent paper, and in his previous one dealing with non-linear differential equations with constant coefficients, Dr. Flake has presented a very straightforward and powerful procedure for treating non-linear systems. There are several comments, however, that I wish to make clear concerning the applicability of the technique.

As has been pointed out, the very serious drawback of the Volterra series representation is the possible divergence of the series; that is, the differential equation may have a perfectly well-defined solution, but the series itself will diverge. For example, the following differential equation with a constant input,  $c$ ,

$$\frac{dy}{dt} + y + ky^2 = x(t) = c$$

has a finite output for  $t \rightarrow \infty$  when

$$-\frac{1}{4} \leq kc$$

but the Volterra series yields a finite solution when

$$-\frac{1}{4} \leq kc \leq \frac{1}{4}$$

Another important point to note is that only those kernels,  $h_n$ , will appear for which

$$n = g + (g-1)k, \quad k = -1, 0, 1, 2, \dots$$

Then, when one is trying to experimentally determine the kernels, it would be quite helpful to know which kernels would not be expected to occur and consequently a great saving of time and effort would be obtained. Moreover, for the case of no initial conditions, then the first-order kernel will be the same as the ordinary impulse response of the non-linear system,  $h(t)$ ,

$$h_1(t) = h(t)$$

However, when the initial conditions are not zero then this may not be true.

Lastly, much of Dr. Flake's work on non-linear differential equations with constant coefficients has been extended to the discrete case and a large class of non-linear difference equations can be solved; the extension to discrete systems with time-varying coefficients should also be possible.

R. H. FLAKE, *in reply*

I thank Dr. Alper for his comments and would add that I have had an opportunity to read some of his recent work applying the Volterra series to the analysis of non-linear discrete systems. I think it would be interesting to investigate how well some of the classical design procedures for linear discrete systems will work for non-linear discrete systems, since there are strong similarities between the analysis of non-linear discrete systems using the Volterra series and multidimensional  $z$  transforms and the usual linear discrete system theory.

R. KULIKOWSKI, *Polish Academy of Sciences, Warszawa ne. Koszykowa 75/18, Poland*

I should like to show, in connection with the very interesting paper of Dr. Flake, an important application of the Volterra series to the non-linear plant identification.

For the sake of simplicity let us assume that the unknown plant can be approximated by the operator

$$y = A(x) = \sum_{i=1}^n \frac{1}{i!} \int_0^t k_i(t-\tau) x^i(\tau) d\tau$$

where  $k_i(t)$  are the unknown continuous functions of time.

Let us also assume that it is possible to determine experimentally (at least approximately) the so-called weak differentials of the plant:

$$\begin{aligned} dA[x, h_1] &= \lim_{\gamma \rightarrow 0} \frac{1}{\gamma} \{A[X + \gamma h_1] - A[X]\} \\ &= \frac{d}{d\gamma} A[x + \gamma h_1]_{(\gamma=0)} \end{aligned}$$

$$\begin{aligned} d^2 A[x, h_1, h_2] &= \lim_{\gamma \rightarrow 0} \frac{1}{\gamma} \{dA[x + \gamma h_2, h_1] - dA[x, h_1]\} \\ &\dots = \dots \end{aligned}$$

$$\begin{aligned} d^n A[x, h_1, h_2 \dots h_n] &= \lim_{\gamma \rightarrow 0} \frac{1}{\gamma} \{d^{n-1} A[x + \gamma h_n \dots h_1] \\ &\quad - d^{n-1} A[x, h_1, \dots h_{n-1}]\} \\ &= \frac{d}{d\gamma} \{d^{n-1} A[x + \gamma h_n, h_1 \dots h_{n-1}]\}_{(\gamma=0)} \end{aligned}$$

This can be done by observation of the output of the plant for the given inputs:  $x(t) + \gamma h_i(t)$ ,  $i = 1, 2, \dots, n$ , where  $\gamma$  is a sufficiently small number. Since

$$dA[0, 1(t)] = \int_0^t k_1(\tau) d\tau$$

$$d^n A[0, 1(t), \dots, 1(t)] = \int_0^t k_n(\tau) d\tau$$

where  $1(t)$  is the step function, we can find experimentally the functions

$$\varphi_i(t) = \int_0^t k_i(\tau) d\tau$$

and then

$$k_i(t) = \varphi'_i(t), \quad i = 1, 2, \dots, n$$

Some more general results concerning this method of non-linear identification were described in a paper in press.

R. H. FLAKE, *in reply*

I thank Professor Kulikowski for his comments on the problem of non-linear identification and I look forward to reading his paper on the subject.

V. S. PUGACHEV, *Institute of Automatics and Telemechanics of the Academy of Science of U.S.S.R., Moscow, I-53, Kalanchevskaya, 15a, U.S.S.R.*

The systems representable by truncated Volterra series considered in Dr. Flake's very interesting paper are the special cases of systems reducible to linear from the statistical point of view. Hence the methods of the statistical theory of linear systems can be applied to solve the problems of statistical analysis and synthesis of such systems. Owing to this fact the characterization of non-linear systems by means of truncated Volterra series has been previously used in problems of statistical studies of non-linear systems. Interesting results were obtained in such a way by Dr. Kulikowsky in Poland and Kichatov in the U.S.S.R. Other applications of the statistical theory of linear systems for the statistical studies of various types of systems reducible to linear have been developed by Dr. Lubbock in England and Dr. Prasad in India.

R. FLAKE, *in reply*

I thank Professor Pugachev for his comments on the use of the Volterra series in the statistical theory of non-linear systems. My first contact with the concept of the Volterra series was through reading Professor Wiener's book on non-linear problems in random theory.

L. P. SHARMA, *Department of Chemical Engineering, University of Virginia, U.S.A.*

The evaluation of the kernels in the Volterra series may offer serious problems, either in difficulty or in accuracy. The first-order kernel must be known, preferably in closed form; if a closed form is not possible, an approximation is necessary. Since an approximate first-order kernel will be used to find the higher-order kernels, the error propagated may grow to large magnitudes in some cases. If the series converges well (as it certainly does in the example cited), the error

should remain tolerable and the solutions will be good. I should like to hear Dr. Flake's comments on the seriousness of these problems.

R. H. FLAKE, *in reply*

When the linear portion of the non-linear differential equation has constant coefficients, then Mr. Sharma's comments do not apply since the exact closed form expression for the first-order kernel may be obtained by routine calculation. However, the more general situation (15), is, I believe, clearly indicated in the paragraph following equation (30) of the paper.

The problem of finding the approximate impulse response of linear differential equations with variable coefficients has been discussed in many papers before and is not the subject of this paper. A survey of these techniques and an investigation of the errors resulting when they are applied in calculating approximate Volterra series solutions for non-linear differential equations with variable coefficients would, I believe, be a worthy subject for further investigation.

# The Applicability of Quasi-linear Methods to Non-linear Feedback Systems with Random Inputs

H. W. SMITH

## Summary

An approximate method is derived for assessing the importance of non-linear effects in feedback systems with Gaussian random inputs, subject to the restriction that the non-linearity be of the zero-memory type. It is shown that the method affords an indication of the applicability of quasi-linear techniques to any specific problem. The choice of a quasi-linear method from among those so far suggested and the class of problem in which it gives accurate results are reconsidered.

## Sommaire

On a établi une méthode approchée pour évaluer l'importance des effets nonlinéaires dans les systèmes à réaction avec des entrées aléatoires, gaussiennes sous la restriction que l'organe nonlinéaire soit du type instantané. On a démontré que la méthode donne une indication de l'applicabilité des techniques quasi-linéaires à tout problème spécifique. Le choix d'une méthode quasi-linéaire parmi celles qu'on a suggérées jusqu'à présent, et le genre de problème dans lequel la méthode donne des résultats précis, sont discutés.

## Zusammenfassung

Eine Näherungsmethode zur Abschätzung des Einflusses von nicht-linearen Effekten in Regelsystemen mit regellosen Eingangsgrößen, die einer Gaußschen Verteilung genügen, wird abgeleitet; die Nicht-linearität ist als verzögerungsfrei vorausgesetzt. Es zeigt sich, daß die beschriebene Methode einen Hinweis auf die Anwendung der Quasi-Linearisierung für jedes spezielle Problem gibt. Die Wahl eines Verfahrens zur Quasi-Linearisierung aus den bisher vorgeschlagenen Methoden und die Gruppe der Probleme, für welche genaue Ergebnisse erzielbar sind, werden betrachtet.

## Introduction

The analysis of non-linear feedback systems with random inputs presents considerable difficulty, and no completely satisfactory method of dealing with such systems now exists. The most easily applied method of analysis is the quasi-linear technique of Booton<sup>1</sup>. In this technique a single-valued zero-memory non-linear device within a feedback loop is replaced by a linear amplifier whose gain is chosen so as to minimize the mean square of the difference between the output of the non-linear device and the output of the linear model which replaces it. The gain of the linear model is a function of the amplitude distribution of the random signal at the input to the non-linearity. If this is assumed to be Gaussian of known mean square value, the linear model is known; linear analysis then allows the determination of the mean square system input of specified spectral characteristics which corresponds to the assumed input to the non-linearity. By repetition, a curve of mean square input against mean square error is found.

It is apparent that this method neglects a component of the output of the actual non-linear element. Kazakov<sup>2</sup>, Axelby<sup>3</sup> and

Pupkov<sup>4</sup> have proposed somewhat more complex linear models which purport to account for this component; in both these models, the assumption of gaussian distribution at the input to the non-linear device is retained.

The statistical properties implied by the three models are:

$$\text{Booton: } \overline{x(t) y'(t+\tau)} = \overline{x(t) y(t+\tau)} \quad (1)$$

$$\text{Kazakov/Axelby: } \overline{[y'(t)]^2} = \overline{[y(t)]^2} \quad (2)$$

$$\text{Pupkov: } \overline{y'(t) y'(t+\tau)} = \overline{y(t) y(t+\tau)} \quad (3)$$

where  $x(t)$  is the input and  $y(t)$  the output of the non-linear device, and  $y'(t)$  the corresponding output of the model. The bar denotes the ensemble average.

It is not possible, in the present state of knowledge, to choose with confidence the best of these methods, nor can the range of applicability of any of them be determined. This paper establishes a basis for choice, gives a method of assessing the accuracy of quasi-linear methods in any specific application, and considers briefly the type of problem to which such methods can effectively be applied.

## The Functional Representation of Non-linear Systems

The functional representation of non-linear systems of Wiener<sup>5</sup> is a powerful analytical method which rests on a firm theoretical basis. However, as Zames<sup>6</sup> has shown, its direct application to feedback systems leads to descriptions of signals which occur within the system in the form of infinite functional series, and serious convergence difficulties arise. At present, therefore, the representation can only be applied to zero-memory non-linear systems which are sufficiently close to linear to allow the convergence of the functional series to be demonstrated; in practice, the restriction is somewhat more stringent, as convergence must be sufficiently rapid to allow solutions to be obtained without excessive labour.

Quasi-linear methods, although not so far justified on theoretical grounds, yield results of moderate accuracy for systems containing violent non-linearities under many conditions of practical interest. The distinctive feature of quasi-linear methods is their consideration of the relations between statistical descriptions of signals, rather than between signals themselves. This suggests that a functional representation in terms of signal statistics might be more informative than the direct representation in terms of signals. The author has shown elsewhere<sup>7</sup> that, if a non-linear feedback system is stable, and if the input to it is

a Gaussian random signal, then: (a) the product of the mean values of any two signals appearing within the system can be expressed as an infinite power series with input-dependent coefficients and the mean value of the input as variable, and (b) the zero-mean second-order cross-correlation function between any two such signals can be expressed as an infinite functional series with input-dependent kernels and the input auto-correlation function as variable. This relation allows the results of previous investigations of the functional representation to be applied here.

### Application to a Zero-mean System

Consider the elementary system shown in Figure 1, which consists of a non-linear element  $N$  and a linear element  $G$  in a simple feedback loop. Let the system be subjected to an input signal  $x(t)$  having zero mean.

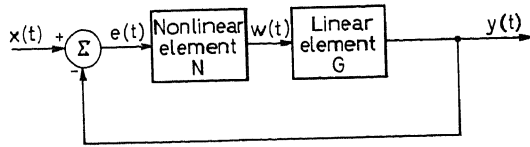


Figure 1. Elementary non-linear feedback system

To form a quasi-functional series describing the statistics of the error signal  $e(t)$ , one must first determine the effect of linear and non-linear operations on statistical quantities. If  $G$  is a linear operator such that

$$y(t) = G * w(t) = \int_{-\infty}^{\infty} h(\tau) w(t-\tau) d\tau \quad (4)$$

then

$$\phi_{wy}(\tau) = \int_{-\infty}^{\infty} h(T) \phi_{ww}(\tau-T) dT = G * \phi_{ww}(\tau) \quad (5)$$

$$\phi_{yw}(\tau) = \int_{-\infty}^{\infty} h(T) \phi_{ww}(\tau+T) dT = \bar{G} * \phi_{ww}(\tau) \quad (6)$$

$$\begin{aligned} \phi_{yy}(\tau) &= \int_{-\infty}^{\infty} h(T_1) \int_{-\infty}^{\infty} h(T_2) \phi_{ww}(\tau+T_1-T_2) dT_1 dT_2 \\ &= \bar{G} * G * \phi_{ww}(\tau) \end{aligned} \quad (7)$$

Also if  $N$  is a zero-memory non-linear operator such that

$$w(t) = N * e(t) = f(e(t)) = \sum_{k=0}^{\infty} C_k (e(t))^k \quad (8)$$

and if  $e(t)$  is Gaussian, then it has been shown<sup>8</sup> that power series relations hold between input and output correlation functions. Using the notation of reference (7),

$$\phi_{ew}(\tau) = \phi_{we}(\tau) = K_1 \phi_{ee}(\tau) \quad (9)$$

$$\phi_{ww}(\tau) = \sum_{k=0}^{\infty} K_k^2 [\phi_{ee}(\tau)]^k = \sum_{k=0}^{\infty} K_k^2 P_k * \phi_{ee}(\tau) \quad (10)$$

$$K_k = \frac{1}{\sigma^k (2\pi \cdot k!)^{\frac{1}{2}}} \int_{-\infty}^{\infty} f(\bar{x} + \sigma \xi) H_k(\xi) \exp\left(-\frac{\xi^2}{2}\right) d\xi \quad (11)$$

$$H_k(\xi) = \exp\left(\frac{\xi^2}{2}\right) \cdot (-1)^k \frac{d^k}{d\xi^k} \exp\left(-\frac{\xi^2}{2}\right) \quad (12)$$

and  $\bar{x}$ ,  $\sigma$  are the mean and standard deviation respectively of the input signal  $e(t)$ .  $K_1$  is Booton's 'equivalent gain'.

For the present purpose, a symmetric non-linearity is assumed, so that all signals within the system of Figure 1 have zero mean. Since

$$x(t) = e(t) + y(t) \quad (13)$$

$$\phi_{xx}(\tau) = \phi_{ee}(\tau) + \phi_{ey}(\tau) + \phi_{ye}(\tau) + \phi_{yy}(\tau) \quad (14)$$

Applying the results (4) to (12), one obtains the operator equation

$$\begin{aligned} \phi_{xx}(\tau) &= (1 + K_1 G) * (1 + K_1 \bar{G}) * \phi_{ee}(\tau) \\ &+ \sum_{\substack{k=3 \\ k \text{ odd}}}^{\infty} \bar{G} * G * P_k * \phi_{ee}(\tau) \end{aligned} \quad (15)$$

This non-linear integral equation must be solved for  $\tau = 0$  if the mean square error and input signals are to be related. However, since it is a quasi-functional equation, a series solution for  $\phi_{ee}(\tau)$  in the form

$$\phi_{ee}(\tau) = E_1 * \phi_{xx}(\tau) + E_3 * \phi_{xx}(\tau) + \dots \quad (16)$$

where  $E_1$ ,  $E_3$ , ... are general functional operators, may, as Zames<sup>6</sup> has shown, be divergent. In fact, investigation of this series solution shows<sup>7</sup>, that, at  $\tau = 0$ , the solution is an alternating infinite series, and that the ratio of successive terms involves, among other factors, the ratio  $K_3/K_1$ . If this ratio is large, the functional solution will fail to converge; this failure of convergence does in fact occur for systems of practical interest.

### An Approximate Solution Method

An approximate solution of the integral equation can be obtained by replacing the terms of the form  $P_k * \phi_{ee}(\tau)$  by white noise approximations  $2 N_k^2 \delta(\tau)$ , where  $N_k^2$  is a constant so chosen as to obtain the best match to the exact terms over the range where they affect the solution. For clarity, this will be discussed with reference to the third-order  $P_3$ . The extension will be clear. In Figure 1,

$$\begin{aligned} \phi_{ee}(\tau) &= (1 + K_1 \bar{G})^{-1} * (1 + K_1 G)^{-1} * \\ &[\phi_{xx}(\tau) - \bar{G} * G * K_3^2 P_3 * \phi_{ee}(\tau)] \end{aligned} \quad (17)$$

approximately. The equivalent linear closed-loop transfer operator is now defined as

$$T = K_1 G * (1 + K_1 \bar{G})^{-1} \quad (18)$$

Rewriting (17) with the aid of (18)

$$\begin{aligned} K_1^2 \phi_{ee}(\tau) &= \bar{G}^{-1} * G^{-1} * T * T * \phi_{xx}(\tau) \\ &- \bar{T} * T * K_3^2 P_3 * \phi_{ee}(\tau) \end{aligned} \quad (19)$$

and, Fourier transforming,

$$K_1^2 \Phi_{ee}(s) = \frac{T(-s)T(s)}{G(-s)G(s)} \Phi_{xx}(s) - T(-s)T(s) \Phi_{ee}^{(3)}(s) \quad (20)$$

where

$$\Phi_{ee}^{(3)}(s) = K_3^2 \int_{-\infty}^{\infty} [\phi_{ee}(\tau)]^3 \exp(-s\tau) d\tau \quad (21)$$

Now for all practical control systems,  $T(s)$  is a band-limited transfer function such that

$$T(-j\omega)T(j\omega) \approx \text{constant}, \quad |\omega| < |\omega_c| \quad (22)$$

$$T(-j\omega)T(j\omega) \approx k\omega^{-2n}, \quad n \geq 1, |\omega| \geq |\omega_c| \quad (23)$$

where  $\omega_c$  is an angular frequency to be called the equivalent linear cut-off frequency of the system. Thus, if the white noise approximation holds over the frequency range  $|\omega| \leq |\omega_c|$ , little error will result.

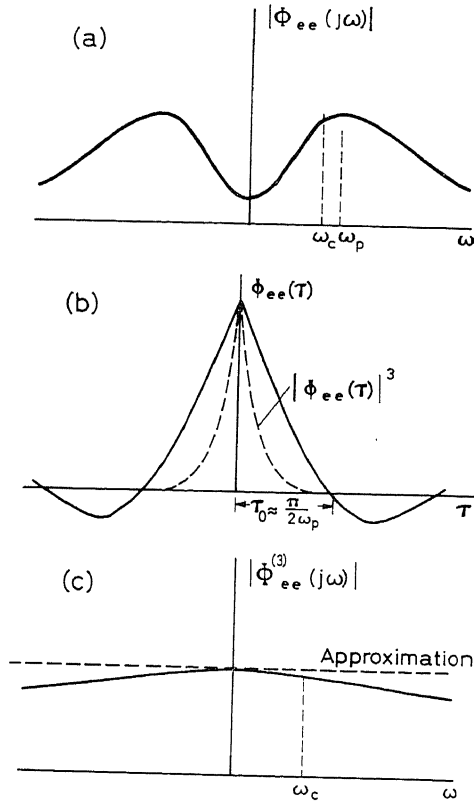


Figure 2. (a) Typical error spectrum within feedback system  
(b) Typical error auto-correlation function  
(c) Typical third-order distortion spectrum

If  $\Phi_{xx}(s)$  represents an input which does not have peaks in its frequency spectrum, and if the system defined by  $T(s)$  does not exhibit serious resonant effects, then  $\Phi_{ee}(s)$  is a smoothly varying function free from line-like spectral components. Also, since it is generally true that feedback systems exhibit little error for inputs at frequencies below their cut-off frequencies, much of the energy in the spectrum of  $e(t)$  lies in the frequency range beyond  $\omega_c$ . Thus, in many systems,  $\Phi_{ee}(s)$  has the form shown in Figure 2 (a).

The corresponding auto-correlation function  $\phi_{ee}(\tau)$  is sketched in Figure 2 (b). Note that (a) all loops of the function

are small in amplitude compared to the first, by virtue of the broad peaking of  $\Phi_{ee}(s)$ , and (b) the first loop of the function extends over a time range inversely proportional to the peaking frequency  $\omega_p$ .

When the operation

$$P_3 * \phi_{ee}(\tau) = [\phi_{ee}(\tau)]^3 \quad (24)$$

is performed, the resultant function has the form shown in Figure 2 (b), and the corresponding power density spectrum is shown in Figure 2 (c). This spectrum is essentially flat for  $|\omega| \leq |\omega_c|$ , and may be approximated over that range by a white noise having the spectrum

$$\Phi_{ee}^{(3)}(s) \approx \Phi_{ee}^{(3)}(0) \quad (25)$$

Using this approximation, the integral eqn (17) can be solved.

### Numerical Example

To illustrate the solution procedure, consider the system of Figure 1, with

$$G(s) = \frac{1}{s} \quad (26)$$

$$\Phi_{xx}(s) = \frac{2B^2}{s(-s)} \quad (27)$$

and

$$N * e = e^3 \quad (28)$$

Then

$$\Phi_{ee}(s) = \frac{2B^2}{(s+K_1)(-s+K_1)} - \frac{1}{(s+K_1)(-s+K_1)} \Phi_{ee}^{(3)}(s) \quad (29)$$

First replace  $\Phi_{ee}^{(3)}(s)$  by the white noise  $2N^2$ , where  $N$  is a constant as yet undetermined.

$$\Phi_{ee}(s) \approx \frac{2(B^2 - N^2)}{(s+K_1)(-s+K_1)} \quad (30)$$

Having found  $\Phi_{ee}(s)$ , determine  $\Phi_{ee}^{(3)}(s)$  by carrying out the operation of eqn (21). The mathematical techniques of George<sup>9</sup> frequently ease the labour of this step. In this case,

$$\phi_{ee}(\tau) = \frac{B^2 - N^2}{K_1} \exp(-K_1|\tau|) \quad (31)$$

$$[\phi_{ee}(\tau)]^3 = \left(\frac{B^2 - N^2}{K_1}\right)^3 \exp(-3K_1|\tau|) \quad (32)$$

$$\Phi_{ee}^{(3)}(s) = K_3^2 \left(\frac{B^2 - N^2}{K_1}\right)^3 \cdot \frac{6K_1}{(s+3K_1)(-s+3K_1)} \quad (33)$$

Now set

$$2N^2 = \Phi_{ee}^{(3)}(0) \quad (34)$$

yielding

$$N^2 = \frac{K_3^2}{3K_1} \left(\frac{B^2 - N^2}{K_1}\right)^3 \quad (35)$$

It is now possible to determine  $\overline{e^2(t)}$ , using Parseval's integral and eqn (33), to obtain

$$\overline{e^2(t)} = \frac{B^2 - N^2}{K_1} \quad (36)$$



Up to this point, the solution has been perfectly general.  $K_1$  and  $K_3$  are now determined using (11), (12) and (28).

$$K_1 = \frac{1}{\sigma(2\pi)^{\frac{1}{2}}} \int_{-\infty}^{\infty} (\sigma\xi)^3 \cdot \xi \exp\left(-\frac{\xi^2}{2}\right) d\xi = 3\sigma^2 = 3\overline{e^2(t)} \quad (37)$$

$$K_3 = \frac{1}{\sigma^3(12\pi)^{\frac{1}{2}}} \int_{-\infty}^{\infty} (\sigma\xi)^3 (\xi^3 - 3\xi) \exp\left(-\frac{\xi^2}{2}\right) d\xi = \sqrt{6} \quad (38)$$

To obtain the relation between  $\overline{e^2(t)}$  and the magnitude of  $B$ , (35) and (36) must be solved simultaneously. This problem reduces to determining the real roots of a single cubic equation. In the example at hand

$$N^2 = \frac{6}{9\overline{e^2(t)}} \cdot [\overline{e^2(t)}]^3 = \frac{2}{3} [\overline{e^2(t)}]^2 \quad (39)$$

and

$$B^2 = K_1 \overline{e^2(t)} + N^2 = 3 \cdot 67 [\overline{e^2(t)}]^2 \quad (40)$$

Thus, finally

$$\overline{e^2(t)} = 0 \cdot 522 B \quad (41)$$

In the more general case, such a simple analytic solution is not possible, and numerical calculation for each value of  $\overline{e^2(t)}$  assumed is necessary.

It will be seen from this example that a solution is obtained if, for all positive values of  $\overline{e^2(t)}$  assumed, a real positive value of mean square input results from the solution of the cubic equation. It is possible to show<sup>7</sup>, by a simple *reductio ad absurdum* proof, that a valid solution is always possible using the approximation made above. It may further be shown that, if the spectra of distortion terms arising in a functional series solution are approximated by white noise in a manner similar to that used above, the functional series solution and the approximate solution are identical whenever the functional series converges.

In addition to guaranteeing a valid (though approximate) solution to the non-linear integral equation for system error, the method illustrated above lends itself well to computation. All terms of the functional series which result from the presence of a third-order term in the non-linear characteristic are accounted for by the solution of a single cubic equation, whereas a functional series solution (if slowly convergent) requires the computation of many spectral terms, followed by the solution of a high-order algebraic equation to obtain the input corresponding to a postulated error. Only in the case of weakly non-linear systems is the labour of computation comparable.

### The Gaussian Assumption

In deriving the error equation, the assumption was made that the input signal to the non-linear element is Gaussian. However, it has been shown<sup>10, 11</sup> that, in cases where the amplitude distribution of this signal can be exactly determined, this assumption does not hold. The value of the technique described above thus lies in its estimation of the importance of non-linear distortion rather than its improved accuracy; whenever the correction to the quasi-linear solution resulting from consideration of distortion effects is large, the Gaussian assumption fails. Thus the method affords a rapid indication of the applicability of quasi-linear methods in any specific instance.

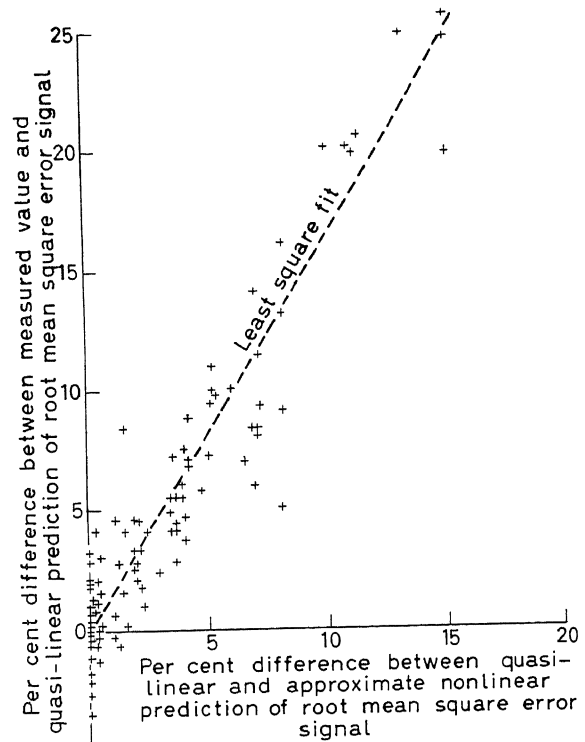


Figure 3. Relation between predicted and measured errors in simulated non-linear systems

To demonstrate the validity of the method, non-linear systems of first and second order containing cubic, relay, and saturation non-linearities have been simulated with random input signals of varying bandwidths applied<sup>7</sup>. A total of 108 simultaneous measurements of mean square input, mean square error, and amplitude distribution of the error signal were made. It was found that, whenever third-order distortion effects (as determined by the approximate solution method) were large, quasi-linear analysis gave results which were appreciably in error (Figure 3). Moreover, where large inaccuracies occurred, the measured distribution of error signal was markedly non-Gaussian (Figure 4).

It may therefore be said that, whenever the effects of non-linear distortion are of appreciable importance, any method of analysis based on the assumption that the signal at the input to the non-linear element in a feedback system is Gaussian will fail to give accurate results. It follows, therefore, that there is no advantage in applying quasi-linear methods more complex than that of Booton<sup>1</sup> in instances where quasi-linear methods are applicable. This is equivalent to ignoring the non-linear term in eqn (15) above. There is some reason to believe<sup>7</sup> that, when the errors of this simple quasi-linear analysis are large, the methods proposed by Axelby<sup>3</sup> and Pupkov<sup>4</sup> are subject to even larger errors.

### The Applicability of Quasi-linear Analysis

The best available simple method of analysis for feedback systems containing a single zero-memory non-linearity, and subjected to a Gaussian random input, is that of Booton (and its extension to non-zero mean systems by Somerville and

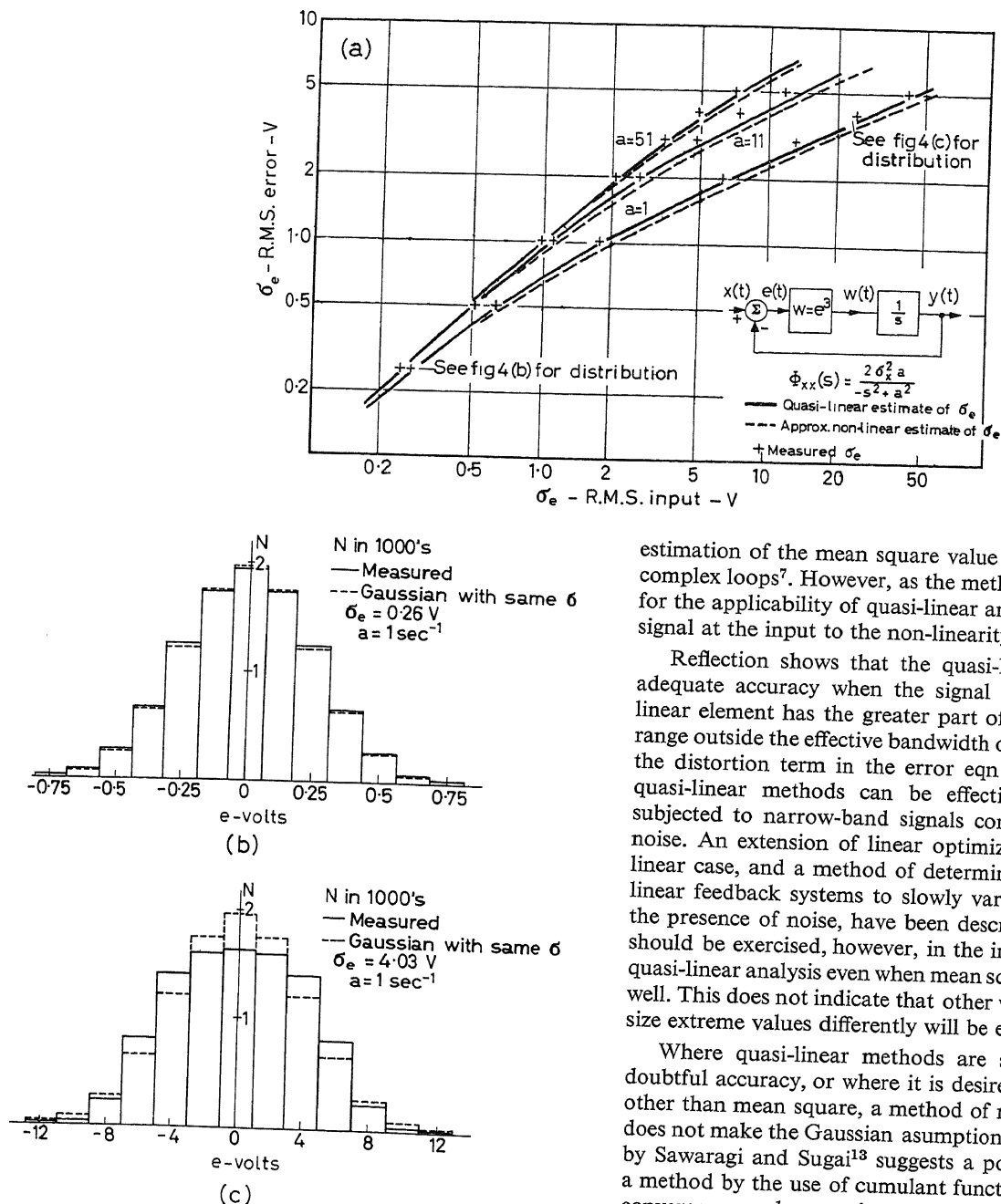


Figure 4. (a) Input-error curves for a typical simulated system  
 (b) Distribution histogram of 10,000 samples of error signal  
 (c) Distribution histogram of 10,000 samples of error signal

Atherton<sup>12</sup>). The method of analysis presented above allows the applicability of the quasi-linear method to any specific case to be determined. When, however, the predicted effect of the non-linear distortion neglected in quasi-linear analysis is large, the Gaussian assumption fails, and no method dependent upon it can give accurate results.

This paper deals with the estimation of mean square error signal in a simple loop. The method can be extended to allow

estimation of the mean square value of any signal within more complex loops<sup>7</sup>. However, as the method proposed is only a test for the applicability of quasi-linear analysis, examination of the signal at the input to the non-linearity will usually suffice.

Reflection shows that the quasi-linear method will afford adequate accuracy when the signal at the input to the non-linear element has the greater part of its power in a frequency range outside the effective bandwidth of the system. In this case, the distortion term in the error eqn (19) will be small. Thus, quasi-linear methods can be effectively applied to systems subjected to narrow-band signals contaminated by wide-band noise. An extension of linear optimization theory to the non-linear case, and a method of determining the response of non-linear feedback systems to slowly varying systematic inputs in the presence of noise, have been described elsewhere<sup>7</sup>. Caution should be exercised, however, in the interpretation of results of quasi-linear analysis even when mean square signals are predicted well. This does not indicate that other weightings which emphasize extreme values differently will be equally well predicted.

Where quasi-linear methods are shown by test to be of doubtful accuracy, or where it is desired to use error weighting other than mean square, a method of non-linear analysis which does not make the Gaussian assumption is required. Recent work by Sawaragi and Sugai<sup>13</sup> suggests a possible approach to such a method by the use of cumulant functions. However, the same convergence and approximation problems apply to this method, since they are in fact inherent in the functional description of the system, and much remains to be done before a practicable engineering method of analysis can result.

## References

- 1 BOORON, R. C. The analysis of nonlinear control systems with random inputs. *Proc. Symp. Nonlinear Circuit Analysis, Polytech. Inst. Brooklyn*, 1953; Nonlinear control systems with random inputs, *Trans. Inst. Rad. Engrs Trans. Prof. Group Circuit Theory* CT-1 (March 1954)
- 2 KAZAKOV, I. E. Approximate probability analysis of the operational precision of essentially nonlinear feedback control systems. (Russian). *Automat. Telemekh., Moscow* (1955)

- <sup>3</sup> AXELBY, G. Random noise with bias signals in nonlinear devices. *Inst. Rad. Engrs Trans. Prof. Group Automat. Control* AC-4, 2 (November 1959)
- <sup>4</sup> PUPKOV, K. A. Method of investigating the accuracy of essentially nonlinear automatic control systems by means of equivalent transfer functions. (Russian). *Automat. Telemekh., Moscow* 21, 2 (February 1960)
- <sup>5</sup> WIENER, N. *Nonlinear Problems in Random Theory*. 1958. New York; Technology Press and Wiley
- <sup>6</sup> ZAMES, G. Nonlinear operators for system analysis. *Sc.D. Thesis, Massachusetts Inst. Technol.* 1960
- <sup>7</sup> SMITH, H. W. Analysis of nonlinear feedback systems with random inputs. *Sc.D. Thesis, Massachusetts Inst. Technol.* 1961
- <sup>8</sup> BARRETT, J. F. and COALES, J. F. Introduction to analysis of non-linear control systems with random inputs. *Proc. Instn elect. Engrs. Monog.* 154M, 103 C (March 1956)
- <sup>9</sup> GEORGE, D. A. Continuous nonlinear systems. *Sc.D. Thesis, Massachusetts Inst. Technol.* 1959
- <sup>10</sup> CHUANG, K. and KAZDA, L. F. A study of nonlinear systems with random inputs. *Trans. Amer. Inst. elect. Engrs* 78 Pt II 42 (May 1959)
- <sup>11</sup> WISHNER, R. P. On Markov processes in control systems. *Ph. D. Thesis, Univ. Illinois.* 1960
- <sup>12</sup> SOMERVILLE, M. J. and ATHERTON, D. P. Multigain representation for a single-valued non-linearity with several inputs, and the evaluation of their equivalent gains by a cursor method. *Proc. Instn elect. Engrs Monog.* 309M, 105 C (July 1958)
- <sup>13</sup> SAWARAGI, Y. and SUGAI, N. Accuracy considerations of the equivalent linearization technique for the analysis of a non-linear control system with a Gaussian random input. (English). *Mem. Fac. Engng Kyoto* 23 Pt III (July 1961)

## DISCUSSION

C. H. P. BROOKES, *Department of Engineering Science, Oxford University, Oxford, England*

The problem of deciding on the accuracy of Booton's quasi-linear approximation for feedback systems containing non-linear elements can also be considered in this way. It is easy to extend this method to obtain information about the size of a simple non-linearity present in the system<sup>1</sup>.

For a linear system with input  $x(t)$  and output  $y(t)$ ,

$$\Phi_{xy}(s) = H(s) \Phi_{xx}(s)$$

and

$$\Phi_{yy}(s) = |H(s)|^2 \Phi_{xx}(s)$$

where  $H(s)$  is the transfer function and  $\Phi_{xx}(s)$  etc. the power spectra.

Eliminating  $H(s)$  from these two equations a spectral ratio  $R(s)$  can be obtained which is always unity.

$$R(s) = \frac{\Phi_{xx}(s) \cdot \Phi_{yy}(s)}{|\Phi_{xy}(s)|^2} = 1 \quad (1)$$

Consider a system such as in *Figure 1* of Dr. Smith's paper with  $H(s)$  as the transfer function of the linear element. The output spectrum may be expressed as a power series expansion,

$$\Phi_{yy}(s) = \sum_{n=1}^{\infty} a_n^2 |H(s)|^2 \Phi_{Nee}^{(n)}(s)$$

where  $a_n$  are the coefficients described by Barrett and Coales (Reference 8 of the paper) and  $\Phi_{Nee}(s)$  is the Fourier transform of the  $n$ th power of the normalized error autocorrelation function.

If the non-linearity equivalent gain is  $K$  and the error R.M.S. value is  $\sigma_e$  then,

$$a_1 = \sigma_e K$$

and higher order terms represent the distortion components caused by the non-linearity.

In addition, quasi-linear analysis gives

$$\Phi_{ey}(s) = KH(s) \Phi_{ee}(s)$$

and from (1)

$$R(s) = 1 + \frac{a_2^2 \Phi_{Nee}^{(2)}(s) + a_3^2 \Phi_{Nee}^{(3)}(s) + \dots}{K^2 \Phi_{ee}(s)} \quad (2)$$

$$= 1 + \delta(s) \quad (3)$$

The value of  $R(s)$  given by (2) will indicate whether or not a quasi-linear analysis of the system is justified because the component of the

non-linearity output which has been ignored will show up as a positive quantity added to the spectral ratio. This also applies to more complex systems with any type of non-linearity.

For a simple type of non-linear element the coefficients  $a_n$  in the output spectrum expansion can be calculated and often this expansion converges quite rapidly. Then the value of  $\delta(s)$  obtained from (3) will enable an estimate of the size of the non-linearity to be made.

In the case of saturation, the even coefficients  $a_2, a_4$ , etc. disappear and

$$\delta(s) \cdot \Phi_{ee}(s) = \frac{a_3^2}{K^2} \Phi_{Nee}^{(3)}(s) + \frac{a_5^2}{K^2} \Phi_{Nee}^{(5)}(s)$$

approximately.

The coefficients  $a_3^2/K^2$  and  $a_5^2/K^2$  can be computed in terms of the ratio  $\sigma_e/h$  where  $h$  is the saturation level. Using these values and experimental results for  $\delta(s)$ ,  $\Phi_{ee}(s)$ ,  $\Phi_{Nee}^{(3)}(s)$  and  $\Phi_{Nee}^{(5)}(s)$  a solution for  $\sigma_e/h$  may be found. It is not necessary to know the value of  $H(s)$ .

The restriction that the input to the non-linearity must have an approximately Gaussian distribution still applies and, of course, the presence of noise in the output will introduce an error by adding to the value of  $\delta(s)$ .

The spectral ratio used here is the inverse of the system coherence discussed by Goodman<sup>2</sup>.

## References

- <sup>1</sup> BROOKES, C. H. P. Evaluation of non-linear systems with random inputs. *M. Eng. Sc. Thesis*, University of Sydney, 1963
- <sup>2</sup> GOODMAN, N. R., KATZ, S., KRAMER, B. and KUO, M. T. Frequency response from stationary noise. *Technometrics*. 2, No. 3, May 1961

W. M. WONHAM, *Research Institute for Advanced Studies, 7212 Bel-lona Avenue, Baltimore 12, Md, U.S.A.*

The author has provided a useful empirical rule for estimating the accuracy of R.M.S. error calculations based on Booton's linearization method. For the author's numerical example, the *exact* value of  $\sqrt{e^2}$  can be found after solving the (equilibrium) Fokker-Planck equation, and is  $0.681/\sqrt{B}$ . Booton's method yields  $0.759/\sqrt{B}$  and the author's method  $0.724/\sqrt{B}$ . These values seem roughly consistent with the graph (*Figure 3*) although it is not quite clear how the percentage differences in *Figure 3* have been defined.

Y. SAWARAGI, *Faculty of Engineering, Kyoto University, Kyoto, Japan*

Since a full variety of the applicability depending on both non-linearities and dynamic characteristics of linear elements within a feedback control system exists, it is necessary to acquire more quantitative

results in due consideration of various combinations of them. As we consider more various types of combinations of non-linearities and linear elements, the procedure presented here becomes more complicated. In what manner will you develop your further study circumventing the difficulty of numerical evaluation?

On the other hand it is very useful to directly evaluate the difference between the probability distribution of the system input and the corresponding output. We have already studied this problem by the method of solving the associated Fokker-Planck equation with the non-linear control equation. I hope that parallel references of our results, concerning the evaluation of the response probability distribution may be helpful to explore your field, for example technical reports of the Engineering Institute of Kyoto University Report Nos. 68 and 79.

G. S. AXELBY, *Westinghouse Electric Corporation, 211 Coronet Drive, North Linthicum, Md., U.S.A.*

Dr. Smith has presented some very useful and interesting results with respect to the amplitude distribution of noise in systems with non-linear elements. I think that more papers of this nature are needed to compare computed and measured data and to obtain more knowledge about the performance of non-linear systems subjected to noisy inputs. In fact, for most systems, this type of input must be considered if realistic results are to be obtained. I would like to ask the author if, in the course of his research, he has extended the work which was stated in Reference 3 of his paper. Actually, I believe that the author may have misinterpreted the results and purpose of this work to which he refers several times. It was the purpose of the reference to develop an engineering technique of calculating the R.M.S. and mean value of an error in a closed-loop system with a non-linear element. This approximate technique was based on a graphical representation of the amplitude distribution. The formula of eqn (2) of Dr. Smith's paper was not used directly, or implied; instead the mean and R.M.S. values of the non-linear element were calculated from the first and second moments of the output amplitude distribution as if they were the centre of gravity and radius of gyration of the amplitude distribution. Thus the method was not restricted to a Gaussian distribution as implied in the paper. In fact, the use of moments for the cases considered gave results more nearly resembling exact computer results than did other methods of calculating the R.M.S. value including that of Booton.

An example was given in Reference 3 where a Gaussian input distribution was assumed to exist with the realization, from previous work, that this was an approximation. The data in Dr. Smith's paper indicates how this assumption may be modified to make a more accurate approximation. However, the work in Reference 3 was concerned with the mechanism of a deterministic input as well as a random input, and it was demonstrated with a graphical solution that the error in a simple first-order system with a saturating element is dependent on the magnitude of the input noise and on the type of input. For example, it was shown that the error would grow without bound if the noise input became too large with respect to the magnitude of a ramp input for which the system would normally have a bounded error.

Dr. Smith indicates in his paper that he has extended his work to include the response of non-linear feedback systems to slowly varying systematic inputs in the presence of noise. I assume that this may also be an extension of Reference 3, and I would like to know if Dr. Smith would comment on this extension of his work possibly pointing out some phenomena of interest.

J. C. WEST, *Electrical Engineering Department, Queen's University of Belfast, Ireland*

I cannot agree with the conclusions of Dr. Smith when he says there is no advantage in using methods more complex than that of Dr. Booton. Too much emphasis is placed on the importance of the amplitude probability distribution. It is of more value to attempt to evaluate the

frequency spectrum of the error, especially where distortion produces power at low frequencies in the working band of the system. In Dr. Smith's method he makes the assumption that the distortion has a white noise spectrum. This is a good approximation for the system's working band and the determination of its magnitude is of extreme importance. It seems to me unimportant that the error does not turn out to be Gaussian. Contrary to the author, I consider his results in *Figure 4* to show that a Gaussian approximation is a good engineering strategy. In fact, the input signal in practice is no better than an approximation.

A Booton linearization leads to an error spectrum approaching zero at zero frequency and this is contrary to practical results. The low frequency distortions as we showed in 1958<sup>1</sup> is probably the most important requirement of the designer.

#### Reference

- <sup>1</sup> WEST, J. C., DOUCE, J. L. and LEARY, B. G. Frequency spectrum distortion of random signals in non-linear feedback systems. *Proc. I.E.E.* 105c, 7 (1958) 88

G. S. AXELBY

In response to Dr. West's remarks, it would seem that the importance of the amplitude or the frequency distribution of the noise input to a non-linearity depends upon the character of the non-linearity and the system in which it is used. For example, knowledge of the amplitude distribution of a signal is most important if it is to be applied to a relay or to a device with a dead zone, because if the R.M.S. value of the signal is sufficiently small with respect to the dead zone, the relay would essentially remain inoperative regardless of its frequency distribution.

H. W. SMITH, *in reply*

Before replying to Professor Sawaragi's question, I should like to acknowledge the thorough work of his group at Kyoto University on many problems arising in non-linear systems with random inputs: the paper<sup>1,2</sup> referred to is the most recent of a series of valuable contributions.

As regards computational difficulty, I feel that, in general, too little heed is paid in this field to the development of good notations and computation techniques. I have found the operator algebra and transform technique developed in my more lengthy work, which extends the work of George<sup>3</sup>, to be much more convenient than manipulation of convolution integrals. Also presented in that work is a numerical approximation method for calculating the constants  $K_k$  defined in eqn (11). This method allows rapid evaluation of the constants either by digital machine methods or by using tabulated functions and a desk calculator. The most laborious part of the calculation is, in high-order systems, the evaluation of the mean square error using Parseval's integral, a problem also encountered in linear systems. When, however, fourth-order statistics are considered, as suggested in my last paragraph, the computational difficulties may become considerable.

Mr. Brookes' method for the experimental evaluation of the type of system considered here is valid, since  $\varphi_{ee}(s)$  is experimentally known. However, even though his spectral ratio is formally independent of the forward path transfer function  $H(s)$ , the low-pass nature of this element in many systems may make experimental measurements of sufficient accuracy difficult. If my contention, that important non-linear effects are accompanied by a change of amplitude distribution of the signal  $e(t)$ , is accepted, a more sensitive test for essential non-linearity within the controlled plant might be the measurement of the amplitude probability distribution at its input. Both methods are, of course, sensitive to the actual location of the non-linearity within the plant, and require somewhat long measurement times; data reduction is, however, considerably easier if distribution is used as the criterion.

I apologize to Mr. Axelby if, in compressing the material presented, I have inadvertently given an erroneous impression of his work. As he states, the method of linearization he used is applicable to any input distribution, not just the Gaussian. However, I believe I am correct in saying that the mean and second moment equivalence properties resulting from his method of calculation are identical to those resulting from Kazakov's first method<sup>2</sup>, and that his work does imply (although it does not explicitly state) my eqn (2).

The method of analysis of systems with both systematic inputs and wideband noise referred to in the paper is a generalization of the work of Somerville and Atherton<sup>12</sup> and Axelby<sup>3</sup>.

For such systems, it is shown<sup>7</sup> that the single system may be reduced to two independent systems: (a) a system incorporating a non-linearity different from that in the actual system, with the noise-free systematic signal as input, and (b) a linear system with quasi-time-varying gain and an input of noise alone. Solution of the first yields mean values of signals within the system, while solution of the second yields variances. A suitable rapid calculation technique similar to that mentioned earlier is also derived.

In verifying the validity of the method experimentally, the increase in error to ramp inputs remarked upon by Mr. Axelby was predicted and found.

The direct solution of the Fokker-Planck equation for the system used as an example is, as Dr. Wonham says, possible. (This is why it was chosen in the first instance.) The figures he quotes and the theoretically correct amplitude probability density function of  $e(t)$ , will be found in the dissertation which this paper partially summarizes<sup>7</sup>. In defining percentage differences, differences between the stated quantities

were taken in the sense indicated, i.e. first quantity minus second, and normalized with respect to the quasi-linear predicted value.

With regard to Professor West's remarks, I feel that, to some extent at least, we are making much the same statement from different points of view. Effects on the frequency spectrum and amplitude distribution of error caused by non-linear distortion signal components go hand in hand. However, it is not difficult to show that (if the Gaussian assumption is made) these distortion components, although uncorrelated with  $e(t)$  in Figure 1, are correlated with  $x(t)$ . Thus any appreciable distortion component in  $y(t)$  may, and from my data does, have a marked effect on distribution. In my opinion, the resulting change of equivalent gain is a more important source of inaccuracy than neglect of some spectral components.

'Good' engineering accuracy is also a matter of opinion, and, because of the logarithmic scale, Figure 4 may be somewhat misleading. As indicated in Figure 3, errors of 25 per cent were found in some systems simulated. My concern was to determine under what conditions the accuracy of analysis is comparable to the accuracy of carefully made measurements of mean square value (say 5 to 10 per cent). I agree that under many circumstances, input spectral data is not so accurately known; however, surely this strengthens the case for using the simplest available method under such circumstances, so long as it is capable of prediction within the accuracy of the data.

In conclusion, it is well to remember that, as Dr. Wonham points out, the quasi-linear methods dealt with here are still largely empirically based and that interpretation of results will remain a matter in which difference in opinion is not only possible, but to be expected, until a better theoretical basis for the methods can be found.

# A Digital Procedure for the Study of Non-linear Systems for Random Processes

T. PRASAD and V. P. SINHA

## Summary

This paper discusses the applicability of digital techniques to various optimization problems and not only demonstrates the utility of these techniques, but confirms that digital computers may be used for the study of plant behaviour.

Results discussed in the early part of the paper show that analytical methods for the classes of problems dealt with involve mathematically intractable expressions; the analogue computational procedures also do not yield satisfactory results.

The explicit digital relationships and the corresponding systems and structures developed in the paper overcome these difficulties. The digital methods are expected to prove most useful because of the availability of high-speed digital computers capable of handling increasingly large amounts of data. Ideas for further investigations are included.

## Sommaire

Ce rapport étudie la possibilité d'appliquer les techniques du calcul numérique automatique à la résolution des problèmes variés d'optimisation, et met en évidence non seulement l'utilité de ces techniques, mais également l'aptitude d'un calculateur numérique à simuler le comportement d'une installation.

Les résultats présentés dans la première partie de ce rapport montrent que les méthodes analytiques conduisent dans certaines classes de problèmes à des expressions mathématiques qui ne peuvent être traitées; les méthodes du calcul analogique ne donnent pas, elles non plus, des résultats satisfaisants.

Ces difficultés sont surmontées au moyen des relations numériques explicites et des systèmes correspondants exposés dans ce rapport. Les méthodes numériques ouvrent des perspectives intéressantes grâce à l'aptitude des calculateurs numériques à grande vitesse à traiter un nombre toujours croissant d'informations numériques. Certaines suggestions sont présentées pour de plus amples études dans ce domaine.

## Zusammenfassung

Eine beträchtliche Anzahl von theoretischen Arbeiten und experimentellen Untersuchungen wurde durchgeführt, um die Zufallsprozesse (regellose Vorgänge) im Hinblick auf Probleme der automatischen Systeme und der Regelungstechnik zu klassifizieren sowie die Untersuchung von nichtlinearen Systemen bei regellosen Vorgängen durchzuführen. Die hierbei angewandten analytischen oder Analogrechenverfahren führen zu simultanen Integralgleichungen, deren Lösungen sehr mühselig und zeitraubend sind. Dieser Beitrag gibt einen kurzen Überblick über die wichtigsten bisherigen Arbeiten auf diesem Gebiet und empfiehlt die Verwendung der von Prasad entwickelten digitalen Analyse für die Untersuchung und Optimierung einer Klasse nichtlinearer Systeme. Einige digitale Systeme mit Nichtlinearitäten erster Ordnung, die auf diesen digitalen Verfahren beruhen sowie einige zugehörige theoretische und experimentelle Gesichtspunkte zur weiteren Untersuchung werden angegeben. Es ist denkbar, daß sich derartige Verfahren nicht nur zur Untersuchung nichtlinearer Systeme erster Ordnung eignen, sondern auch auf eine systematische Untersuchung von Erscheinungen bei regellosen Signalen führen.

## Study of Random Processes

The classification of random processes on the basis of their statistical characteristics and evaluation of mathematically tractable analytical expressions for their correlation properties or probability distribution functions have baffled a number of research workers over the last few years. A fair amount of progress has nevertheless been recorded. Some important contributions have been made by Busgang<sup>8</sup>, Brown<sup>7</sup>, Nuttal<sup>18, 19</sup> and Lubbock<sup>14</sup>. Busgang has shown that for two Gaussian signals, the cross-correlation function taken after one of them has undergone non-linear amplitude distortion is identical with the cross-correlation function taken before the distortion. He has supplemented his theoretical deductions with a number of examples dealing with various types of non-linear distortion. Nuttal has extended the work of Busgang and has defined a general class of separable processes. The random processes belonging to this class are characterized by second-order probability density functions that satisfy the equation

$$p(x_1, t_1; x_2, t_2) = p_1(x_1, t_1) p_2(x_2, t_2) \times \sum_{n=0}^{\infty} a_n(t_1, t_2) \theta_n^{(1)}(x_1, t_1) \theta_n^{(2)}(x_2, t_2) \quad (1)$$

where

$$p(x_1, t_1) = \int_{-\infty}^{\infty} p(x_1, t_1; x_2, t_2) dx_2$$

$$p(x_2, t_2) = \int_{-\infty}^{\infty} p(x_1, t_1; x_2, t_2) dx_1$$

$$\int_{-\infty}^{\infty} p_1(x_1, t_1) \theta_m^{(1)} \theta_n^{(1)} dx_1 = \delta_{mn}$$

and

$$\int_{-\infty}^{\infty} p_2(x_2, t_2) \theta_m^{(2)} \theta_n^{(2)} dx_2 = \delta_{mn}$$

The correlation functions of such processes possess several invariance properties which may be utilized to advantage<sup>18</sup>. Lubbock's work deals with a still more general class of semi-separable processes. Two processes  $x_1$  and  $x_2$  belonging to this class are semi-separable with respect to each other if:

$$g_1(x_1; t_1, t_2) = g_1(x_1, t_1) g_2(t_1, t_2) \quad \text{for } t_1 \geq t_2 \quad (2)$$

where

$$g_1(x_1; t_1, t_2) = \int_{-\infty}^{\infty} (x_2 - m) p_2(x_1, x_2; t_1, t_2) dx_2 \quad (3)$$

$m$  being the ensemble average and  $p_2(x_1, x_2; t_1, t_2)$  being the joint probability density function of  $x_1$  and  $x_2$ .

The invariance properties of these are the same as those of the separable processes of Nuttal over the restricted range  $t_1 \leq t_2$  or  $t_2 \leq t_1$ .

## Study of Non-linear Systems

Almost simultaneously with the classification work, the study of non-linear systems for random processes, particularly with reference to automation and control problems, has led to considerable theoretical work and experimental investigations. Singleton<sup>22</sup> was perhaps the first to study non-linear filters for random processes. His work embodies: (a) characterization of the input in terms of its past values by samples at the instants  $t, t - T, \dots, t - (N - 1)T$ , and (b) specification of a system operator  $F$ , the input-output relation being

$$y(t) = F[x(t), x(t - T), \dots, x(t - \overline{N - 1} T)] \quad (4)$$

For determining the optimum filter, higher order correlation functions of the input and higher order cross-correlation functions between the input and output are required.

The works of Booton, Mathews and Siefert<sup>4, 5</sup> outline a method of evaluating non-linear characteristics of servo-mechanisms from their responses to Gaussian signals. The non-linearity discussed by them is characterized by

$$X_{AR} = f(X_{AI}) \quad (5)$$

where  $X_{AR}$  and  $X_{AI}$  denote the instantaneous amplitudes of the response and input respectively. The response may be separated into a quasi-linear and a harmonic term as

$$X_{AR} = K_{eq} X_{AI}(t) + X_H(t) \quad (6)$$

The optimum value of the describing function  $K_{eq}$  for least mean square error is given by

$$K_{eq} = \frac{\int_{-\infty}^{\infty} X_{AI} f(X_{AI}) p_1(X_{AI}) dX_{AI}}{\int_{-\infty}^{\infty} X_{AI} p_1(X_{AI}) dX_{AI}} \quad (7)$$

A theory for the experimental determination of optimum time-invariant non-linear systems was developed by Bose<sup>6</sup>. The experimental procedure outlined by him for the evaluation of the coefficients of the optimum filters involves the use of a large number of gate circuits, averaging devices etc.

Zadeh<sup>21</sup> studied non-linear systems in terms of classes  $\pi_n$ ,  $n = 1, 2, 3, \dots$ , consisting of all two poles for which the input-output relationship may be expressed in the form of an  $n$ -fold integral,

$$y(t) = \int_0^\infty \dots \int_0^\infty K[x(t - \tau_1), \dots, x(t - \tau_n); \tau_1, \dots, \tau_n] d\tau_1, \dots, d\tau_n \quad (8)$$

where the characteristic function  $K$  is any real function of the variables  $x(t - \tau_1), \dots, x(t - \tau_n)$ .

Following Zadeh, Lubbock<sup>15</sup> used the Barret Lampard expansion<sup>3</sup> to study the filters of a subclass of  $\pi_1$  more comprehensively. His results are in compact and accessible forms, leading to an almost routine procedure for the determination of the optimum weighting functions of this class.

It is evident from the examples of Lubbock that the analytical procedures or analogue computational methods adopted for optimization of even the simplest class of non-linear filters of Zadeh<sup>24</sup> for stationary random processes whose statistical prop-

erties are describable by very simple expressions, lead to a set of simultaneous integral equations whose solutions are very tedious and time consuming.

The non-linear correlation function ordinates and the digital analysis of input-output statistics developed by Prasad<sup>20</sup> for the study and optimization of Zadeh's filters of class  $\pi_1$  overcome the complexities which are encountered when analogue methods are used<sup>20, 21</sup>. The utility of these techniques has been further confirmed by Sinha<sup>23</sup> who has used them in making a comparative study of several types of filters for separating non-Gaussian signals from Gaussian noise in the mean square sense. He obtained a random input by superimposing 1,000 samples of a random Gaussian noise over 1,000 samples of a given non-Gaussian signal. The mean square errors for the types of filters studied by him are reproduced in Table 1 given below. The values reveal that the optimum non-linear filters with storage give a considerably smaller mean square error and one may hope to reduce the error further by increasing the number of channels in the multipath non-linear filters<sup>23</sup>.

Table 1

Serial No.	Type of filter	Mean square error, normalized with respect to that for the linear filter
1	Linear	1.00
2	Instantaneous non-linear power series with the first five terms	1.65
3	Instantaneous non-linear power series with the first ten terms	1.12
4	Storage non-linear filter (two followed by a suitable linear filter)	0.85
5	Storage non-linear filter (three followed by a suitable linear filter)	0.68

It may be mentioned here that the study of the problem carried out by Sinha would have been not only extremely difficult but perhaps impossible without the help of the computational procedures and formulae of Prasad.

## Utilization of Digital Techniques

The scope of application of the digital techniques of Prasad is not limited to the optimization of filters only. These may also be found extremely useful in the approximate analysis of non-linear systems for stochastic processes in general. Four different aspects are dealt with here.

### Equivalent Linear Structures for Non-linear Systems of First Order<sup>21</sup>

The non-linear element of Figure 1 may be approximated by a linear network  $G_L$ . If  $G_{NL}$  is an instantaneous non-linear element, it is easily shown that for the class of separable processes characterized by the equation

$$\phi_{f_x f_x}(\tau) = K \phi_{f_x f_x}(\tau), \tau \geq 0 \quad (9)$$

where  $\phi$  is the staircase correlation function defined by Prasad<sup>20</sup>.

$G_L$  may be a simple attenuator of gain  $A$ . The optimum value of  $A$  is obtained by minimizing the mean square error

$$\bar{\epsilon}^2 = \frac{Lt}{T_0 + T} \frac{1}{2} \int_{-T_0}^{T_0+T} \left\{ \sum_n f(x_n) - A \sum_n x_n \right\}^2 P(t-nT) dt \quad (10)$$

For minimum error,

$$A = \frac{\phi_{fxf}}{\phi_{xx}} \quad (11)$$

In a different situation,  $G_{NL}$  may be a storage non-linear element of the first order (class  $\pi_1$  of Zadeh) as shown in Figure 2.

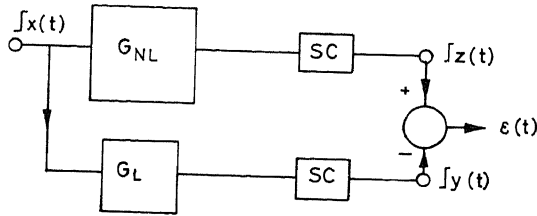


Figure 1. Replacing non-linearities by linear structures

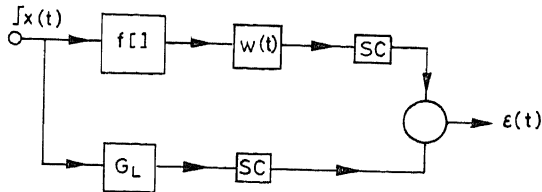


Figure 2

In that case,  $G_L$  may be allowed to be a linear network of impulsive response  $g(t)$ . The corresponding  $P$ -response ordinates  $\gamma_n$  for minimum mean square error are given by the set of simultaneous equations

$$\sum_{r=0}^N \gamma_r \phi_{fxf}(|r-m|T) = \sum_{r=0}^N u_r \phi_{fxf}(|r-m|T) \quad (12)$$

$$m=0, 1, 2, \dots, N$$

The impulsive response function  $g(t)$  may then be obtained from  $\gamma$  by interpolation techniques discussed by Prasad<sup>20</sup>.

Still another case of an instantaneous non-linear element whose approximate structure may be sought for is shown in Figure 3.  $f[]$  is a no-memory non-linear element operating on a set of input functions  $x_0, x_1, x_2, \dots, x_M$ . The approximating structure is a multichannel linear network consisting of  $M$  at-

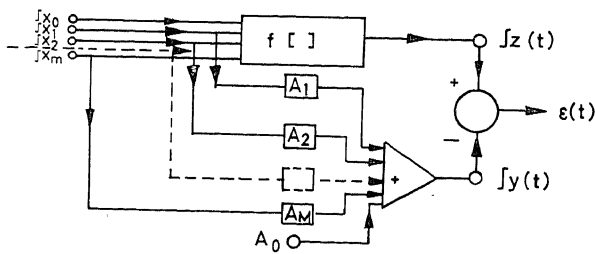


Figure 3. Equivalent gains for several random inputs

tenuators  $A_1, A_2, \dots, A_M$  and a constant term  $A_0$ . The minimizing equations reduce to a set of simultaneous algebraic equations that may be solved for the various  $A$ 's.

#### Approximate Structures for Non-linear Elements Controlling a Plant

Figure 4 shows the schematic of a plant describable by a known linear impulsive response  $g(t)$  controlled by a non-linear element  $G_{NL}$ . It is desired to approximate  $G_{NL}$  by a linear network  $G_L$  such that the mean square error between  $\int z(t)$  and  $\int y(t)$  is minimum.

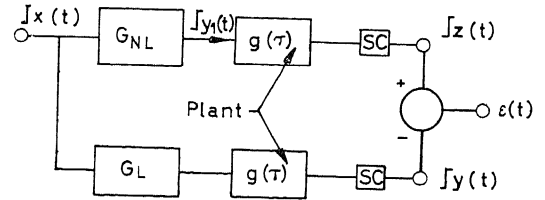


Figure 4. Equivalent structure for non-linear controller of a plant

If  $G_{NL}$  is a no-memory element  $f[]$ ,  $G_L$  may be an attenuator whose gain  $A$  is given by the minimizing equation

$$A \sum_r \sum_s \gamma_r \gamma_s \phi_{fxf}(|r-s|T) = \sum_r \sum_s \gamma_r \gamma_s \phi_{fxf}(|r-s|T) \quad (13)$$

where  $\gamma$  is the  $P$ -response ordinate of  $g(t)$ . The  $g(t)$  itself may then be evaluated from the values of  $\gamma$ .

If  $G_{NL}$  is a storage non-linear device  $\{f[]\}$  followed by  $w(t)$ , a linear memory structure may be chosen as  $G_L$ . Its  $P$ -response ordinate  $\alpha$  for minimum mean square error is given by the simultaneous algebraic equations

$$\sum_{r=0}^N \alpha_r \sum_{j=0}^{\infty} \phi_{fxf}(jT) \phi_{fxf}(|r-s|+j)T$$

$$= \sum_{r=0}^N u_r \sum_{j=0}^{\infty} \phi_{fxf}(jT) \phi_{fxf}(|r-s|+j)T \quad (14)$$

$$s=0, 1, 2, \dots, N$$

where

$$\phi_{fxf}(jT) = \sum_{m=0}^{\infty} \gamma(mT) \gamma(m+jT)$$

#### Equivalent Structures for Non-linear Feedback Systems

Figure 5 depicts the essential features of a unity feedback system having a non-linear element  $G_{NL}$ . A possible equivalent structure (in the mean square sense) is shown in Figure 6 in which  $G_{NL}$  has been replaced by two optimum boxes  $F$  and  $G$ .

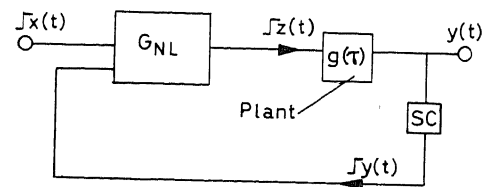


Figure 5. Non-linear feedback staircase system



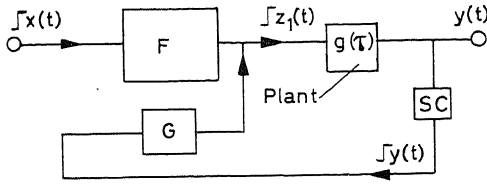


Figure 6. Feedback system equivalent to that of Figure 5

For the same system output in both the cases, the mean square error between the outputs  $\int z(t)$  and  $\int z_1(t)$  is given by

$$\bar{\varepsilon}^2 = \frac{dt}{T_0 + T} \int_{-T_0}^{T_0+T} \{ \int z(t) - \int z_1(t) \}^2 dt \quad (15)$$

If  $G_{NL}$  is an instantaneous non-linear piece  $f[\cdot]$ , an appropriate set of  $F$  and  $G$  may be

$$F = \sum_{i=0}^{M-1} A_i f_i[\cdot] \quad (16)$$

and

$$G = A \delta(t)$$

where  $f_i$  is a known non-linear function,  $A$  and  $A_i$ ,  $i = 0, 1, \dots, (M-1)$ , are unknown constants to be determined. For minimum mean square error, these are given by

$$\sum_{i=0}^{M-1} A_i \phi_{\int y \int f_i}(0) - A \phi_{\int y \int f_r}(0) = \phi_{\int y \int f_r}(0) \quad (17)$$

and

$$\sum_{i=0}^{M-1} A_i \phi_{\int f_i \int f_r}(0) - A \phi_{\int f_r \int f_r}(0) = \phi_{\int f_r \int f_r}(0)$$

for  $r = 0, 1, 2, \dots, M-1$

Other modified structures, linear as well as non-linear, may be obtained on the same lines if  $G_{NL}$  is a no-memory non-linear element.

### Self Optimizing Systems

A somewhat different class of problems in which the staircase techniques are again useful relates to self-optimizing control systems. These form a topic of considerable interest at present. A comprehensive bibliography of research carried out in this field is given in the work of Aseltine *et al.*<sup>2</sup> Some more recent publications are those of Anderson<sup>1</sup>, Kalman<sup>13</sup> and Margolis and Leondes<sup>16</sup>.

Based upon the principles of operation, self-optimizing systems may be broadly classified as learning systems and computing systems.

The following is confined to computing systems only.

The block schematic of an input sensing self-optimizing or adaptive system is shown in Figure 7. The plant  $P$  is a time-invariant structure described by its staircase  $P$ -response  $\int \gamma(t)$ . The quantities

$$\phi_{\gamma\gamma}(jT) = \sum_{k=0}^{\infty} \gamma(kT) \gamma(\overline{k+jT})$$

are assumed to be fixed during the interval the system operates. The output of the plant,  $\int y(t)$ , is to follow a desired output  $\int z(t)$  in the mean square sense; the input  $\int x(t)$  may have slowly

varying statistical characteristics. The plant  $P$  is preceded by a self-optimizing controller,  $SOC$ . It is a varying parameter network described by its impulsive response

$$w(t) = A_0 \delta(t) + \sum_{r=1}^N A_r e^{-K_r t} \quad (18)$$

where  $K_r$  is a fixed time constant and  $A_r$  is a varying gain under the control of the signals received from  $E_3$ . The control unit consisting of  $E_{xx}$ ,  $E_{yy}$ ,  $E_1$ ,  $E_2$ ,  $E_3$  is prescribed to set the gains of the  $SOC$  to such values as maintain minimum mean square error between  $\int z(t)$  and  $\int y(t)$ .

The plant output is

$$\int y(t) = \sum_n \sum_k \sum_r u_r x_{k-r} \gamma_{n-k} P(t-nT) \quad (19)$$

The optimizing equations for the system are

$$\sum_{r=0}^N u_r \alpha_{|r-s|} = \beta_s \quad (20)$$

for  $s = 0, 1, 2, \dots, N$ ;  $|r-s|$  represents the modulus of  $r-s$  where

$$\alpha_K = \sum_{j=0}^N \phi_{\gamma\gamma}(jT) \phi_{\int x \int x}(\overline{k+jT})$$

for all  $k \geq 0$  and

$$\beta_m = \sum_{j=0}^N \gamma(jT) \phi_{\int x \int z}(\overline{m+jT})$$

for all  $m \geq 0$

The  $\alpha$  and  $\beta$  are evaluated by the computing units  $E_1$  and  $E_2$  respectively.  $E_{xx}$ ,  $E_{zz}$  and  $E_{yz}$  supply the various correlation ordinates required for the computation. Eqn (20) is solved for the staircase ordinates  $u$  from which the gains  $A$  of the  $SOC$  may be obtained as discussed earlier. The computer corrector  $E_3$  solves eqns (20) in two stages and finally sets the values of the gains  $A_0, A_1, A_2, \dots, A_N$  accordingly.

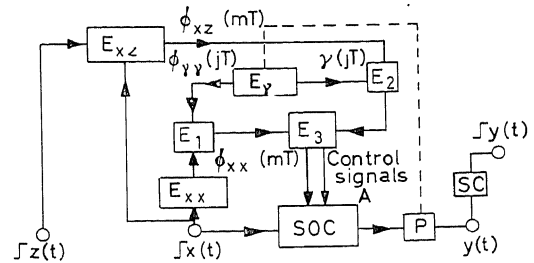


Figure 7. Self-optimizing staircase control system (linear controller)

The plant of Figure 7 has been assumed to be time invariant. The theory of optimization may be easily extended on the same lines to time-varying plants also. In this case, an additional computing mechanism has to be incorporated in the assembly in order to evaluate the varying dynamic characteristics of the plant in terms of  $P$ -response ordinates. This is shown in the sectionalized diagram of Figure 8.

A further generalization results if we also remove the restrictions of structural linearity on the  $SOC$  and the plant. The situation, however, becomes difficult even for non-linear plants of class  $\pi_1$  of Zadeh. Explicit optimizing relationships for the  $SOC$  in this case cannot be derived without an approximate linearization of the characteristics.

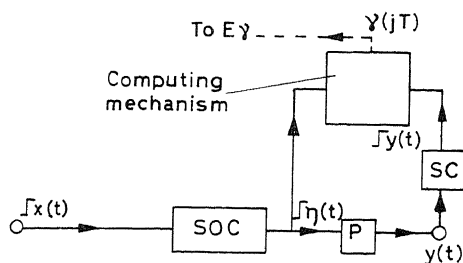


Figure 8. A computing mechanism for plant behaviour

### Conclusions and Remarks

The foregoing discussion throws enough light on the applicability of the digital techniques to various optimization problems. The cases considered<sup>20, 21</sup> not only demonstrate the utility of these techniques but also confirm that digital computers may be confidently used for the study of plant behaviour. The digital techniques may also lead to a systematic study of random phenomena. Further experimental and theoretical researches in this field may perhaps provide adequate foundations for comprehensive study of adaptive and learning mechanisms, problems involving non-stationary stochastic processes and time-invariant structures, linear or non-linear.

The results summarized in the earlier sections of this paper readily show that analytical methods for the classes of problems dealt with involve mathematically intractable expressions; the analogue computational procedures also do not yield satisfactory results. The explicit digital relationships and the corresponding systems and structures discussed subsequently appear to overcome these difficulties. The digital methods are expected to prove most useful because of the availability of high-speed digital computers capable of handling increasingly large amounts of data. A few ideas for further investigations are outlined below.

(a) A major experimental investigation constitutes the study of self-optimizing controllers, linear as well as non-linear, for slowly varying random processes described by auto-correlation functions of the type

$$\phi(\tau) = \rho_1 e^{-\mu_1 |\tau|} + \rho_2 e^{-\mu_2 |\tau|} + \dots + \rho_n e^{-\mu_n |\tau|}$$

in the stationary state. Another problem which may be pointed out is that of studying the systems of first order for non-stationary random staircase inputs by introducing known variations in their stationary time domain records.

(b) The analyses presented<sup>20</sup> are restricted to staircase random functions that are synchronized with respect to a fixed closure time. Experimental and theoretical study of non-linear staircase systems may be carried out for cases in which two closure periods bearing integral fractional ratios to each other are involved. Other first order non-linear systems which contain more than two sampler clamps may likewise be considered for multiple sampling, multirate sampling and mixed sampling<sup>17</sup>.

(c) The linear and non-linear correlation functions that form the basis of the staircase techniques characterize the behaviour of the system only at the sampling instants. A rigorous or systematic approach for the analysis of first order systems in between the hold times is lacking. Perhaps a fictitious delay on the same lines as that proposed by Jury<sup>11</sup> may be introduced to obtain theoretical expressions giving the information between the sampling instants.

(d) The staircase techniques may be applied satisfactorily to practical problems with any desired degree of accuracy if the random functions are well behaved. It is suggested, however, that erratic random processes may be dealt with in a better manner by adopting the periodic sampling procedure as proposed by Jury and Mullin<sup>12</sup>.

(e) The  $p$ -transform of Farman-Farmaian<sup>9, 10</sup> for sampled-data control systems with finite pulse width offers a promising avenue for the study of the sampled non-linear systems of first order. Apparently, any time function multiplied by the interpolatory unit pulse of Farman-Farmaian may be seen to possess discontinuous but periodic properties similar to those of the staircase function. Nothing, however, can be asserted unless comprehensive analytical and experimental work is carried out on this topic.

### References

- ANDERSON, G. W., ASELTINE, J. A., MANCINI, A. R. and SARTURE, C. W. A self-adjusting system for optimum dynamic performance. *Nat. Conv. Rec., Inst. Radio Engrs N.Y.* 4 (1958) 182
- ASELTINE, J. A., MANCINI, A. R. and SARTURE, C. W. A survey of adaptive control system. *Trans. Inst. Radio Engrs N.Y. Automatic control*. December (1958) 102
- BARRET, J. F. and LAMPARD, D. G. An expansion for some second-order probability distributions and its application to noise problems. *Trans. Inst. Radio Engrs N.Y. IT-I*, No. 1, March (1955)
- BOOTON, R. C., JNR. Non-linear control systems with statistical inputs. *Dynamic Analysis and Control Lab., M.I.T., Rep. No. 61*, March (1962)
- BOOTON, R. C. JNR., MATHEWS, M. V. and SIEFERT, W. W. Non-linear servo-mechanisms with random inputs. *Dynamic Analysis and Control Lab., M.I.T., Rep. No. 7*, August (1953)
- BOSE, A. G. A theory of non-linear systems. *Tech. Rep. 309, Res. Lab. Electronics, M.I.T.*, May (1953)
- BROWN, J. L. On a cross-correlation property for stationary random processes. *Trans. Inst. Radio Engrs I.T.-3* (1957) 28
- BUSSGANG, J. J. Cross-correlation functions of amplitude distorted Gaussian signals. *Tech. Rep. No. 216, Res. Lab. Electronics, M.I.T.* March (1952)
- FARMAN-FARMAIAN, G. Analysis of linear sampled-data systems with finite pulse width (open loop). *Trans. Amer. Inst. elect. Engrs* 75, Pt 1 (1956)
- FARMAN-FARMAIAN, G. Analysis of multiple sampler systems with finite pulse width (open loop). *Trans. Amer. Inst. elect. Engrs* 77, Pt II (1958)
- JURY, E. I. *Sampled Data Control Systems*. 1958. New York; Wiley
- JURY, E. I. and MULLIN, F. J. The analysis of sampled data control systems with a periodically time-varying sampling rate. *Trans. Inst. Radio Engrs N.Y. AC-4*, No. 1, May (1959)
- KALMAN, R. F. Design of a self-optimizing control system. *Trans. Amer. Soc. mech. Engrs* 80 (1958) 468
- LUBBOCK, J. K. *On a Class of Semi-separable Processes*. University of Cambridge. Unpublished
- LUBBOCK, J. K. Optimization of a class of non-linear systems. *Automatic and Remote Control*. 1960. London; Butterworths
- MARGOLIS, M. and LEONDES, C. T. A parameter tracking servo for adaptive control systems. *Inst. Radio Engrs, Wescon Conv. Rec.*, Pt IV (1959)
- NISHIMURA, T. and JURY, E. I. Contribution to statistical design of sampled-data control system. *Electronic Res. Lab., University of California Berkeley*

- <sup>18</sup> NUTTAL, A. H. Invariance of correlation functions under non-linear transformations. *Tech. Rep., Electronics Res. Lab., M.I.T.*, April (1957)
- <sup>19</sup> NUTTAL, A. H. Theory and applications of the separable class of random processes. *Tech. Ref., Electronics Res. Lab., M.I.T.* May (1958)
- <sup>20</sup> PRASAD, T. Theory and applications of a class of non-linear staircase systems. *Ph. D. Dissert.* 1960. University of Cambridge
- <sup>21</sup> PRASAD, T. Analysis and optimization of a class of non-linear staircase systems for random processes. *Automation and Remote Control; Proceedings of Moscow Conference.* 1961. London; Butterworths
- <sup>22</sup> SINGLETON, H. E. Theory of non-linear transducers. *Tech. Rep.* No. 60, *Electronics Res. Lab., M.I.T.* August (1950)
- <sup>23</sup> SINHA, V. P. and NARESH, K. Optimum non-linear filters for random processes. *Tech. Rep.* New York University 1961
- <sup>24</sup> ZADEH, L. A. A contribution to the theory of non-linear systems. *J. Franklin Inst.* 255, No. 5 (1953)

## DISCUSSION

E. I. JURY, *University of California, California, U.S.A.*

Statistical study of non-linear discrete systems is drawing much attention and interest at the present time, therefore this paper constitutes a timely and valuable contribution in this field. In a separate study on a slightly different problem we were able to extend the statistical technique to discrete systems and specifically to pulse width modulated feedback systems. This work appeared in April 1962.

Like Dr. Prasad, we found the separability property discussed by Dr. Nuttal to be of much use in our statistical study of linearization.

Extensive experimental study on digital computer simulation of the non-linear discrete systems had verified the validity of this approximation; the error involved in most cases was confined to 5 per cent and this is a rather comforting sign for engineering work. It would be interesting to know if Dr. Prasad has conducted an experimental simulation study of the discrete system and, if so, whether the error of approximation is within engineering approximation.

As a final remark, I mention that the modified  $z$  transform could be used more conveniently than the  $P$  transform. This has been carried out in our work<sup>1</sup>.

### Reference

- <sup>1</sup> GUPTA, G. S., and JURY, E. I. Statistical study of pulse width modulated feedback systems. *J. Franklin Inst.*, April 1962

T. PRASAD, *in reply*

I thank Professor Jury for advocating the utility of digital techniques for the study of non-linear systems for random processes. The mathematical study reported in this paper appears to be important for two reasons; first, because of the fact that the non-linear systems discussed are capable of using more statistical information, and secondly because of the fact that the techniques reported here, and the structures proposed with their corresponding computational formulae, are envisaged to be of practical use to the engineer. Professor Pugachev has already referred to these approximation techniques<sup>1</sup> in his survey paper read at the 2nd Congress, Basle, 1963.

The reply to Professor Jury's question may be given in the affirmative on the basis of the work of Sinha and Naresh (Reference 23 of the paper), outlined. More complicated and practical experiments also can be simulated as suggested here; the engineers and experimenters from other disciplines have to undertake the task. It may be added that all the non-linear digital structures or systems of first-order discussed in this report are stable in the sense of Professor Tsytkin's criterion presented at this Congress<sup>2</sup>.

### References

- <sup>1</sup> PRASAD, T. Analysis and optimization of a class of non-linear staircase systems for random processes. *Automatic and Remote Control (Theory)* 1961. London; Butterworths
- <sup>2</sup> TSYPKIN, Ya. Z. Fundamentals of the theory of non-linear pulse control systems. *Automatic and Remote Control (Theory)* 1964. London; Butterworths: Munich; Oldenbourg

# Time-optimal Systems with Random Noise Disturbances

V.N. NOVOSELTSEV

## Summary

Time-optimal control systems with random noise disturbances are discussed in the paper. It is shown that for a class of control systems the controller programme which provides minimum time of the transient for expected values of system coordinates, assures also the minimum expected value of the transient time. The construction method of switching surfaces to provide the minimum of expected transient time based upon the conventional equation of optimal switching surface, and characteristics of different noisy channels of the system, is given. The examples of the second-order time-optimal systems with noise disturbances are considered. Results of analogue computer simulation of the control systems with plant transfer function in the presence of random noise disturbances applied in various points of the system are given.

## Sommaire

Cette communication a pour objet les systèmes de réglage à temps optimal en présence de perturbations aléatoires constituant un 'bruit'. On y montre que pour une classe de systèmes, le programme de réglage qui, pour atteindre des points de fonctionnement prédéterminés, rend la durée du régime transitoire minimale donne en même temps à ce minimum une valeur égale au minimum de la durée prédite du régime transitoire. On indique, pour la construction des surfaces de commutation, la méthode qui assure le minimum de cette durée prédite, méthode qui est fondée sur l'équation habituelle de la surface de commutation optimale; on donne aussi les caractéristiques de différents canaux 'bruyants' au sein du système. On traite à titre d'exemple les systèmes du second ordre à temps optimal sujets à un 'bruit' perturbateur. On cite enfin les résultats obtenus au cours de la simulation de systèmes de réglage sur calculatrice analogique, en tenant compte de la fonction de transfert de l'installation en présence de perturbations du type 'bruit' appliquées en divers points du système.

## Zusammenfassung

Der Aufsatz behandelt zeitoptimale Regelsysteme mit rauschförmigen Störsignalen. Es zeigt sich, daß für eine Klasse von Regelsystemen die Reglereinstellung, die die kürzeste Übergangszeit für gewünschte Werte der Systemparameter erzeugt, auch mit Sicherheit die kürzeste Übergangszeit für den Ensemblemittelwert ergibt. Das angegebene Konstruktionsverfahren für die Schaltfläche zur Erzielung der kürzesten Übergangszeit beruht auf der üblichen Gleichung für die optimale Schaltfläche. Auch die Eigenschaften verschiedener rauschbehafteter Kanäle des Systems werden behandelt. Zeitoptimale Systeme 2. Ordnung mit rauschförmigen Störgrößen dienen als Beispiele. Die Ergebnisse der Nachbildung am Analogrechner von Regelsystemen mit einer gegebenen Übertragungsfunktion der Strecke und mit an verschiedenen Stellen des Systems einwirkenden rauschförmigen Störgrößen werden verglichen.

## Introduction

This paper examines the problem of optimal control of a plant with constant coefficients, having one input and one output:

$$\dot{x} = f^1(x) + v \cdot f^2(x) \quad (1)$$

Here  $x$  is an  $n$ -dimensional vector which defines the state of the plant,  $v$  is the control signal sent to its input.

Functions  $f^1(x)$  and  $f^2(x)$  are defined and continuous for all  $x$  and continuously differentiable with respect to all coordinates of vector  $x$ :

$$x_i = d^i x / dt^i$$

Equation (1) is linear with respect to  $v$  and non-linear with respect to  $x$ , and is therefore somewhat more general than the equation

$$\dot{x} = Ax + Bu$$

usually considered for the case of a scalar control signal  $u$ .

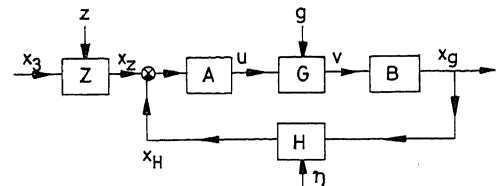


Figure 1

Figure 1 shows a block diagram of the system under consideration in the presence of interference; the following symbols are used:

$A$ —zero-memory controller

$B$ —controlled plant (1)

$H$ —zero-memory plant coordinate metering channel

$G$ —zero-memory control signal-to-plant channel

$Z$ —master-signal channel

$h, g$  and  $z$ —random interference in channels  $H, G$  and  $Z$  respectively with known characteristics

$u$ —control signal (scalar)

$v$ —noise-distorted control signal

$x_d$ —true state of plant

$x_n$ —observed state of plant

$x_z$ —set point of system phase space

$x$ —error vector  $x = x_z - x_d$

The true-state point  $x_d$  has to be shifted into some small vicinity of the set point (origin of the phase space). Then in the optimal system the following equality must be satisfied<sup>1, 2</sup>:

$$E\{T[x^{(0)}]\} = \min$$

The minimal value of  $E\{T(x)\}$  will be denoted by  $T^*(x)$ . Then in the optimal system

$$E\{T[x^{(0)}]\} = T^*[x^{(0)}] \quad (2)$$

Here  $x^{(0)}$  is the initial value of the error vector.

However, considerable difficulties are involved in calculating the system directly from this criterion. It is far simpler to use the criterion of the minimum time of the transient process for the mathematical expectation (the minimum time required to bring the system to a state of statistical non-displacement)<sup>3, 4</sup>. Usually considered for the determination of this time is the relationship  $\xi = E\{x\}$ , which describes the transient processes in some equivalent system without interference. A system in which is provided the minimum time of the transient process for the mathematical expectation  $T(\xi) = \theta(x)$

$$\theta(x) = \min \theta^*(x) \quad (3)$$

will be termed optimal with respect to the criterion  $\theta^*$ . A system in which condition (2) is satisfied, will be termed optimal with respect to the criterion  $T^*$ .

Functions  $T^*(x)$  and  $\theta^*(x)$  are defined for all points of the phase space and  $T^*(x) \geq 0$ ,  $\theta^*(x) \geq 0$  while

$$T^*(x) = \theta^*(x) = 0$$

when, and only when, the point  $x$  lies in the set vicinity of the origin

$$\sum_{i=0}^{n-1} (x_i)^2 \leq \delta^2 \quad (4)$$

Consideration will be given to the state of the control system only at discrete moments of time, as in solving similar problems by the dynamic programming method. For this the small interval of time  $\Delta$  will be introduced and it will be considered that on this interval the values of the control signal and the interference signals remain invariant, but at moments of time  $t = k\Delta$  they change stepwise. Interference on neighbouring intervals will be considered independent.

Control signal  $u$  is constrained with respect to the modulus

$$|u| \leq N \quad (5)$$

To simplify the examination it will be assumed that both  $u$  and  $v$  are quantized in level with a sufficiently small pitch of quantization, and:

$$\begin{aligned} u \in \Omega(u) \quad \Omega(u) &= \{u_1, u_2, \dots, u_r\} \quad r < \infty \\ v \in \Omega(v) \quad \Omega(v) &= \{v_1, v_2, \dots, v_l\} \quad l < \infty \end{aligned}$$

In such a case it is convenient to describe the influence of random interferences in the following way.

Since at the controller  $A$  there arrives only the value of the error  $x$  distorted by the noises along channels  $H$  and  $Z$ , instead of the correct, necessary control signal  $u$  at each moment of time another control signal, generally speaking distinct from  $u_0(x^{(k)})$ , will be chosen. The probability of the choice of control signal  $u$  at a given moment of time, when in fact control signal  $u^0$ , denoted by  $p_{u^0 u}$  is optimal, depends in the general case on the position of the image point  $x$  in the system phase space. Thus, in a relay system this probability strongly depends on distance of the state point from the switching surface. When this distance is great, the probability of error in choice of the control signal is small, but as this distance is reduced even very weak interference can lead to error in the choice of control signal. If one denotes the probability of an event which consists

in the appearance on the output of  $A$  of signal  $u_m$ , whereas the optimal choice would be  $u^0(x) = u_l$  by  $p_{ml}$ , then

$$\|p_{u^0 u}(x)\| = \begin{bmatrix} p_{11}(x) & p_{12}(x) & \dots & p_{1r}(x) \\ p_{21}(x) & p_{22}(x) & \dots & p_{2r}(x) \\ \dots & \dots & \dots & \dots \\ p_{r1}(x) & p_{r2}(x) & \dots & p_{rr}(x) \end{bmatrix} \quad (6)$$

The control signal  $u(x^{(k)}) = u_m$  chosen at the  $k$ th moment by controller  $A$  reaches the plant along the channel with noisy  $G$ , where, under the influence of interference  $g$ , control signal  $u$  becomes control signal  $v$ . The probability of the transformation of  $u_i$  into  $v_j$  will be denoted by  $q_{ij}$ . Then

$$\|q_{uv}\| = \begin{bmatrix} q_{11} & q_{12} & \dots & q_{1l} \\ q_{21} & q_{22} & \dots & q_{2l} \\ \dots & \dots & \dots & \dots \\ q_{r1} & q_{r2} & \dots & q_{rl} \end{bmatrix} \quad (7)$$

Both matrices  $\|p_{u^0 u}(x)\|$  and  $\|q_{uv}\|$  can be joined into one, which will fully describe the action of all the interference upon the system:

$$\|p_{u^0 v}\| = \|p_{u^0 u}(x)\| \|q_{uv}\| \quad (8)$$

Matrix (8) determines for each point of the phase space the probability of the arrival at the input of the plant of control signal  $v_j$ , when  $u_i$  is the optimal signal.

#### Basic Relationship for Time-optimal Systems with Interference

In the optimization of control systems with respect to the criterion  $T^*$  by the dynamic programming method, the following equation can be obtained:

$$T^*[x^{(k)}] = \Delta + \min_{u^{(k)} \in \Omega(u)} T^*[x^{(k+1)}] \quad (9)$$

Here  $x^{(k)}$  and  $u^{(k)}$  denote the values of  $x$  and  $u$  on the  $k$ th interval of time. Equation (9) is the basic relationship in solving such problems<sup>2, 5</sup>. This relationship will be given another form more suitable for the purposes of this paper<sup>6</sup>.

In open form eqn (9) is written as follows:

$$T^*[x^{(k)}] = \Delta + \min_{(m)} \left\{ \sum_{j=1}^l p_{mj}(x^{(k)}) \cdot T^*[x^{(k+1)}]_j \right\} \quad (10)$$

$$m = 1, 2, \dots, r$$

In the latter equation  $\min \{\alpha_m\}$  denotes the minimum of the numbers  $\alpha_m$ , and  $[x^{(k+1)}]_j$  the position of the state point at the  $(k+1)$ th moment of time, provided that at the  $k$ th moment the point was in the position  $x^{(k)}$  and at the plant input there arrived control signal  $v_j$ . Thus the  $m$ th term of the expression in brackets is simply the mathematical expectation of the time of the transient process when control signal  $u_m$  is chosen at point  $x^{(k)}$ . Averaging is performed for all the states of the system at the  $(k+1)$ th moment of time. The probabilities  $p_{mj}$  in (10) are elements of the matrix  $\|p_{u^0 v}\|$ , which is determined by (8).

By introducing the sampling interval  $\Delta$  eqn (1) can be rewritten in the form

$$\begin{aligned} x_i^{(k+1)} &= x_i^{(k)} + \varphi_i^1 + v \cdot \varphi_i^2 \quad i = 0, 1, \dots, n-1 \\ x_n^{(k+1)} &= v^{(k+1)} \end{aligned}$$

where, for brevity, is written  $\Delta \cdot f^1 = \varphi^1$ ;  $\Delta \cdot f^2 = \varphi^2$ . In subsequent operations the relationship to  $x^{(k)}$ , where possible, is dropped. If in the expansion of  $T^*$  into a series, it is possible to limit ourselves (for sufficiently small  $\Delta$ ) to terms of no higher than the first order of smallness, then

$$\begin{aligned} T^*[x^{(k+1)}]_j &\equiv T^*[x_0^{(k+1)}, x_1^{(k+1)}, \dots, x_{n-1}^{(k+1)}]_j \\ &= T^*[x_0^{(k)}, x_1^{(k)}, \dots, x_{n-1}^{(k)}]_j \\ &\quad + \sum_{i=0}^{n-1} \frac{\partial T^*[x^{(k)}]}{\partial x_i} [\varphi_i^1 + v_j \varphi_i^2] \end{aligned} \quad (11)$$

Substituting (11) into each bracketed term in (10), after elementary transforms

$$\begin{aligned} \sum_{j=1}^l p_{mj} T^*[x^{(k+1)}]_j &= T^* + \sum_{i=0}^{n-1} \varphi_i^1 \frac{\partial T^*}{\partial x_i} \\ &\quad + \sum_{i=0}^{n-1} \left\{ \frac{\partial T^*}{\partial x_i} \cdot \varphi_i^2 \cdot \sum_{j=1}^l p_{mj} v_j \right\} \end{aligned} \quad (12)$$

is obtained.

The following notations will be introduced

$$v_m^*(x^{(k)}) \equiv E\{v|u_m(x^{(k)})\} = \sum_{j=1}^l p_{mj}(x^{(k)}) \cdot v_j \quad (13)$$

$$[\xi^{(k+1)}]_m = E\{x^{(k+1)}|x^{(k)}, v_m^*(x^{(k)})\} \quad (14)$$

Here  $v_m^*(x^{(k)})$  is the mathematical expectation of the signal  $v$ , and  $[\xi^{(k+1)}]_m$  the mathematical expectation of the random vector  $x^{(k+1)}$  for the choice of control signal  $u^{(k)} = u_m$  at the point  $x^{(k)}$ . Then, on the basis of (11)–(14), the initial relation (9) can be written in the form:

$$T^*[x^{(k)}] = \Delta + \min_{(m)} \{T^*[\xi^{(k+1)}]_m\} \quad m=1, 2, \dots, r \quad (15)$$

Control signal  $u_m$ , with which the minimum is reached, is the optimal control signal at the point  $x^{(k)}$  of the system phase space with respect to criterion  $T^*$ .

Equation (15) is the basic relation in examination of time-optimal systems with noise present.

*The optimal control action at each moment of time must be so selected that the magnitude of  $T^*[\xi^{(k+1)}]$  at the next step is minimal.*

#### Control Algorithm in a System Optimal with Respect to Criterion $T^*$

Consideration is given to the point  $x^{(k)}$  and the sequence of control actions  $u^{(k)} = u_\alpha$ ,  $u^{(k+1)} = u_\beta$ , ...,  $u^{(k+S)} = u_\sigma$  under the influence of which the state point  $\xi$  shifts consecutively from position  $x^{(k)}$  to positions  $[\xi^{(k+1)}]_\alpha$ ,  $[\xi^{(k+2)}]_{\alpha\beta}$ , ...,  $[\xi^{(k+S)}]_{\alpha\beta\sigma}$ , ...,  $\sigma$ . Let the sequence  $u_\alpha, u_\beta, \dots, u_\sigma$  be selected in such a way that:

(A) The point  $[\xi^{(k+S)}]_{\alpha\beta\sigma}$  lies in the vicinity of the set point (4).

(B) For all sequences  $u^{(k)}, u^{(k+1)}, \dots, u^{(k+d)}$  where  $d < S$ , condition A is not satisfied.

Then the sequence  $u_\alpha, u_\beta, \dots, u_\sigma$  is optimal with respect to the criterion  $\theta^*$ , and  $S$  is the number of steps in the optimal transient process with respect to the criterion  $\theta^*$ . It will be temporarily assumed that there exists and is known a control algorithm

optimal with respect to the criterion  $\theta^*$ , which permits the construction of such a sequence of control signals for any  $x$ .

The recurrent relation (15) will now be applied to the point  $[\xi^{(k+1)}]_m$ :

$$T^*[\xi^{(k+1)}]_m = \Delta + \min_{(m,n)} \{T^*[\xi^{(k+2)}]_{mn}\} \quad n=1, 2, \dots, r \quad (16)$$

Substituting the resultant expression back into (15), one obtains:

$$T^*[x^{(k)}] = 2\Delta + \min_{(m,n)} \{T^*[\xi^{(k+2)}]_{mn}\} \quad (17)$$

$$m=1, 2, \dots, r; n=1, 2, \dots, r$$

similarly

$$T^*[x^{(k)}] = S \cdot \Delta + \min_{(m,n,\dots,l)} \{T^*[\xi^{(k+s)}]_{mn\dots l}\} \quad (18)$$

$$m=1, 2, \dots, r; n=1, 2, \dots, r; l=1, 2, \dots, r$$

One now selects  $m = \alpha, n = \beta, \dots, l = \sigma$ . Then, by virtue of condition A, the second addenda in the right-hand side of (18) vanishes, and, bearing in mind condition B,

$$T^*[x^{(k)}] = S \cdot \Delta \quad (19)$$

can be written.

Thus in the case under consideration the duration of the transient process with respect to criteria  $T^*$  and  $\theta^*$  is identical, and at each point of the phase space the control signals optimal with respect to  $\theta^*$  and  $T^*$  coincide.

*The control algorithm optimal with respect to criterion  $\theta^*$ , ensures the optimality of the system with respect to criterion  $T^*$  as well.*

Consideration is now given to the construction of algorithms optimal with respect to the criterion  $\theta^*$ . The plant studied is linear and is described by the equation

$$x_i = \sum_{j=0}^{n-1} a_{ij} x_j \quad i=0, 1, \dots, n-1 \quad (20)$$

$$x_n = v$$

It will also be assumed that the control algorithm, which ensures time-optimality in the absence of interference, is known and set in the form of a switching surface in an  $n$ -dimensional space:

$$\psi(N, x_0, x_1, \dots, x_{n-1}) = 0 \quad (21)$$

Equation (21) contains in an explicit form the magnitude  $N$  from eqn (5). In the absence of interference the optimal control signal has the form

$$u^0 = -N \operatorname{sign} \psi(N, x) \quad (22)$$

and only the values of  $v = u = \pm N$  reach the plant input.

If in the control system (20)–(21) there are noises, then in place of a system with interference, an equivalent system without interference can be considered, in which instead of the coordinates  $x(t)$  the relationships  $\xi(t) = E\{x(t)\}$  are considered. The optimal control signal in such an interference-free system will be optimal with respect to the criterion  $\theta^*$  in the initial system with interference, and will hence be optimal with respect to the criterion  $T^*$ .

The maximum and minimum values of the signal  $v^*$  which reaches the plant input when  $|u| \leq N$  will be denoted by  $v_M^*$  and  $v_m^*$ .

For the symmetrical matrices (6)-(8)  $v_M^* = -v_m^*$ . For simplicity, the examination will be confined to the case when the signals  $v_M^*$  and  $v_m^*$  are obtained following the selection on the controller of control actions  $u = +N$  and  $u = -N$  respectively.

Introduced here is the coefficient of efficiency of control in a system with interference, which has an obvious sense:

$$\gamma(x) = \frac{v_M^*(x)}{N} = -\frac{v_m^*(x)}{N} \quad (23)$$

It is convenient to examine as an equivalent system a system:

$$x_i = \sum_{j=0}^{n-1} a_{ij} x_j; x_n = u \quad (24)$$

$$i = 0, 1, \dots, n-1$$

and the constraint  $|u| \leq N$  is replaced by  $|u| \leq N \cdot \gamma(x)$ .

The following three cases of the action of interference upon the system are considered:

(1) In the system, interference is present only in channel  $G$ . Here, as follows from (7),  $\gamma = \text{const.}$ , and constraint (5) is replaced by the constraint  $|u| \leq \gamma N$ .

The plant equation remains unchanged. The optimal switching surface has the form

$$\psi(\gamma N, x_0, x_1, \dots, x_{n-1}) = 0 \quad (25)$$

(2) Interference is absent from the channel  $G$ , but the influence of interference  $h$  and  $z$  manifests itself in the appearance of additive noise along the coordinates  $x_0, x_1, \dots, x_{n-1}$ ; at the controller  $A$  there arrive the values

$$\begin{aligned} x_0^* &= x_0 + \eta_0 \\ x_1^* &= x_1 + \eta_1 \\ &\dots \\ x_{n-1}^* &= x_{n-1} + \eta_{n-1} \end{aligned}$$

while the random component is constrained with respect to the modulus:

$$|\eta_i| \leq \eta_i^*, \quad i = 0, 1, \dots, n-1 \quad (26)$$

In this case it can be shown that the optimal control of  $n$ -order plant with real non-positive roots follows the  $n$ -interval theorem and the switching surface has the form:

$$\psi(N, x_0 + \eta_0^* \lambda_0, \dots, x_{n-1} + \eta_{n-1}^* \lambda_{n-1}) = 0 \quad (27)$$

In this equation,  $\lambda_i = \pm 1$ . For the second-order control system  $\lambda_i = -\text{sign } x_i$ ,  $i = 0, 1$ .

Eqn (27) may be obtained by the 'inverse motion' method (see for example ref. 7).

(3) In the system there is present both interference in the channel  $G$  and unconstrained interference  $\eta$ .

In this case it is possible to construct approximately optimal control systems, in which the duration of the transient processes exceeds the minimum possible time by not more than the preset  $\varepsilon$ .

One such system is considered in Example (2).

## Examples

### (1) The Optimal Second-order System with Noise in the Communication Channel

Consideration is given to a control system which has been thoroughly studied for the case of no noise; the block diagram of this is given in Figure 2.

The equation of the optimal switching line of the system without interference has the form:

$$x_0 = -\frac{x_1^2}{2kN} \text{sign } x_1 \quad (28)$$

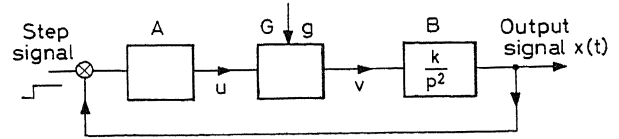


Figure 2

It will be taken that under the influence of interference  $g$  the control signal is able at each moment of time to adopt independently one of the following values

$$v = \begin{cases} a_1 u & \text{with probability } p_1 \\ a_2 u & \text{with probability } p_2 \\ \dots & \dots \\ a_m u & \text{with probability } p_m \end{cases} \quad (29)$$

( $u$  adopts the value  $\pm N$ ). In this case

$$\gamma = \sum_{i=1}^m a_i p_i$$

and in accordance with (25) the equation of the optimal switching line in the system with interference (29) has the form

$$x_0 = -\frac{x_1^2}{2kN \sum_{i=1}^m a_i p_i} \text{sign } x_1 \quad (30)$$

### (2) The Second-order System with Noise in Channel H

The block diagram in Figure 3 is considered. Here in the channel serving for metering the coordinate  $x_1$  the additive interference  $\eta$  is a Gaussian noise with zero mean value:

$$p(\eta) = \frac{1}{\sigma_\eta \sqrt{2\pi}} \exp \left\{ -\frac{\eta^2}{2\sigma_\eta^2} \right\} \quad (31)$$

It is required to ensure time optimality with accuracy of no less than 5 per cent with an aperiodic transient process.

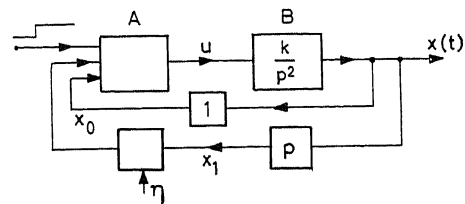


Figure 3

It is known that transient processes without overshoot in the system under review correspond to the switching lines

$$x_0 = -\frac{x_1^2}{akN} \operatorname{sign} x_1 \quad (32)$$

when  $a \leq 2$ , while a 5 per cent extension of the transient process in the processing of step signals is obtained when  $a = 1.650$ .

In a relay control system the coefficient  $\gamma(x)$  is determined by the equation

$$\gamma(x) = 1 - 2p_H(x) \quad (33)$$

where  $p_H(x)$  is the probability of wrong choice of control signal at the point  $x$ :

$$p_H(x) = P\{\operatorname{sign} u^0(x) = -\operatorname{sign} u(x)\}$$

The magnitude of  $\gamma(x)$  rises as the distance  $r_1 = x_{1n} - x_1$  increases, and at some  $r_1^*$  becomes greater than  $0.825 = \frac{1}{2} \cdot 1.650$ . Here  $x_{1n}$  is the coordinate  $x_1$  of a point lying on the switching line and having a coordinate  $x_0$  identical with the point under consideration  $x_{0n} = x_0$ :

$$\psi(x_{0n}, x_{1n}, N) = 0$$

For all the points  $x$ , for which  $r_1 > r_1^*$ , the magnitude of  $\gamma(x)$  will be replaced by  $\gamma^* = 0.825$ . The resultant control system with interference will possess the property, that  $\gamma(x) \leq 0.825$  for all, while for  $r_1 < r_1^*$ ,  $\gamma < 0.825$ .

The reduction of  $\gamma(x)$  when  $r_1 < r_1^*$  stems from the presence of the constrained interference  $\eta^*$  which only manifests itself when  $r_1 < r_1^*$ :

$$|\eta^*| < r_1^* \quad (34)$$

The examination of a system of control of a plant has been arrived at with the equation  $k/p^2$  under the constraint

$$|u| \leq 0.825 N$$

[in such a system the switching-line equation has just the form of (32)], while on the system there acts the constrained interference (34). It only remains to find the magnitude of  $r_1^*$  and substitute it into eqn (27).

From the general formula

$$P\{\alpha < \eta < \beta\} = \frac{1}{2} \left[ \Phi\left(\frac{\beta}{\sigma_\eta \sqrt{2}}\right) - \Phi\left(\frac{\alpha}{\sigma_\eta \sqrt{2}}\right) \right]$$

taking (33) into account, one obtains

$$\gamma^* = \Phi\left(\frac{r_1^*}{\sigma_\eta \sqrt{2}}\right)$$

and for  $\gamma^* = 0.825$  one has  $r_1^* = 1.343 \sigma_\eta$ .

The equation of a switching line which is optimal with accuracy up to 5 per cent has the form:

$$x_0 = -\frac{(x_1 - 1.343 \sigma_\eta \operatorname{sign} x_1)^2}{1.650 kN} \operatorname{sign} x_1 \quad (35)$$

## Experimental Results

Figures 4 and 5 show the graphs of response to step signal of 20 V amplitude in a system controlling a plant  $1/p^2$ ,  $N = 20$  V.

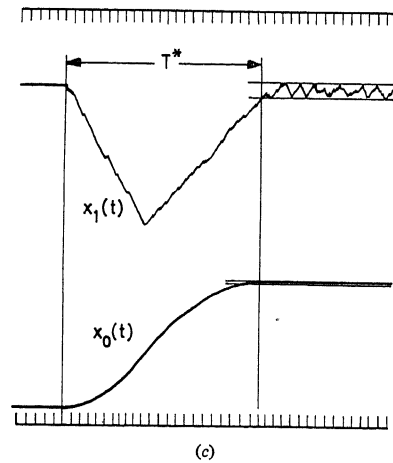
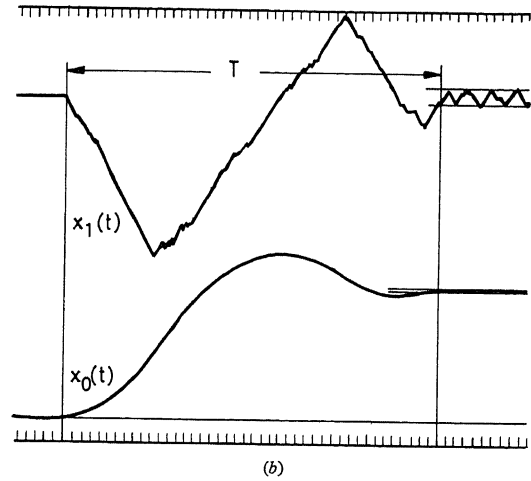
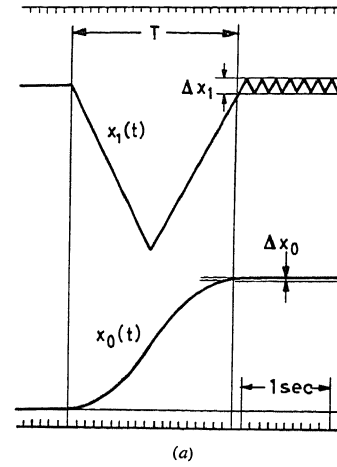


Figure 4

Figure 4 corresponds to Example 1, and in Figures 4(b) and 4(c)  $\gamma = 0.645$ . The switching line equation in Figures 4(a) and 4(b) has the form:

$$x_0 = -\frac{x_1^2}{40} \operatorname{sign} x_1 \quad (36)$$



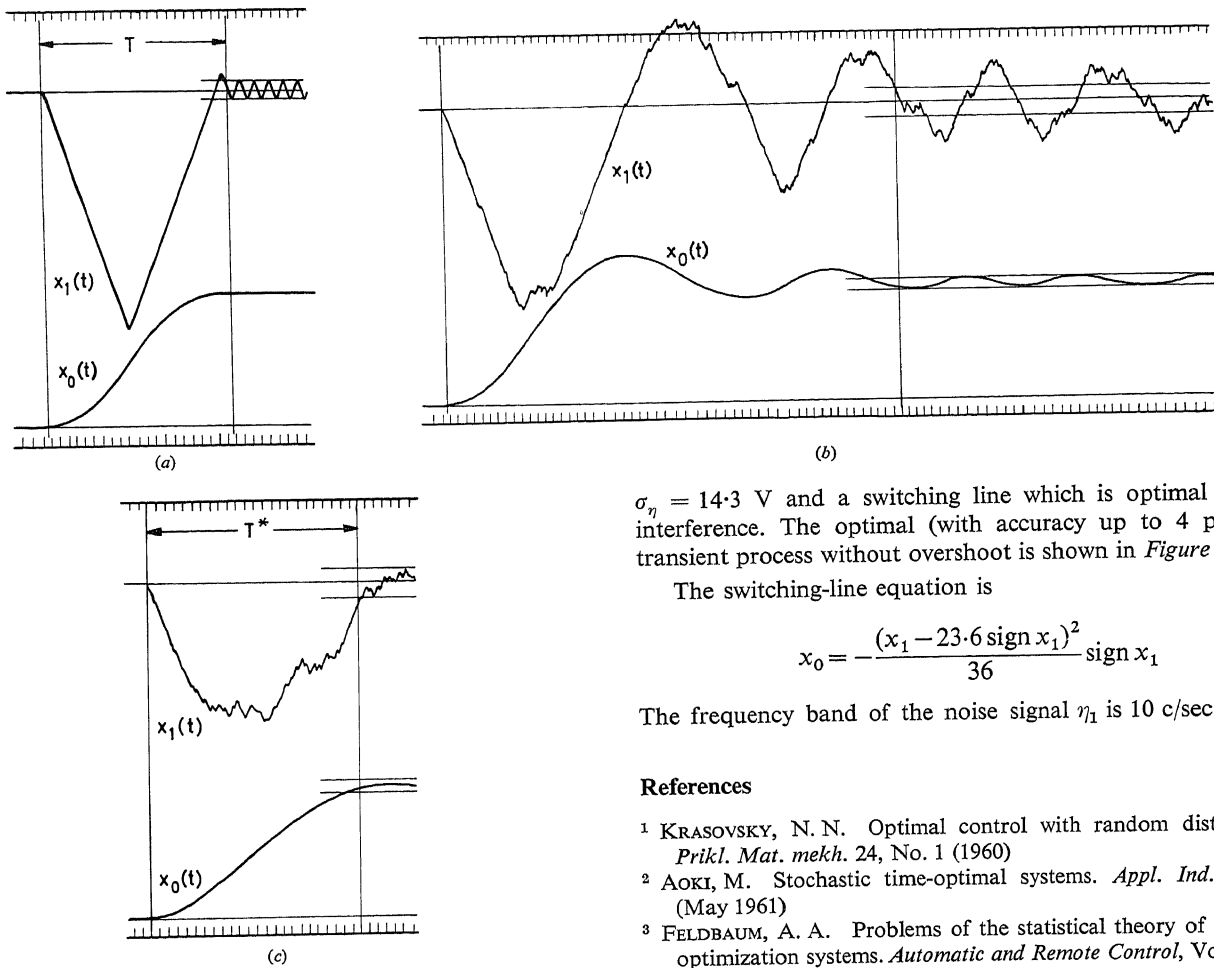


Figure 5

The optimal transient process (Figure 4) is ensured in a system with interference following the choice of the switching line

$$x_0 = -\frac{x_1^2}{25.8} \text{sign } x_1$$

Figures 5(a), (b) and (c) illustrate Example 2 for  $\varepsilon = 4$  per cent,  $\gamma^* = 0.90$ . Figure 5(a) shows the performance of a step signal of amplitude 20 V without interference with switching line (36). Figure 5(b) demonstrates the performance of the signal for

$\sigma_{\eta} = 14.3$  V and a switching line which is optimal without interference. The optimal (with accuracy up to 4 per cent) transient process without overshoot is shown in Figure 5(c).

The switching-line equation is

$$x_0 = -\frac{(x_1 - 23.6 \text{ sign } x_1)^2}{36} \text{sign } x_1$$

The frequency band of the noise signal  $\eta_1$  is 10 c/sec.

## References

- 1 KRASOVSKY, N. N. Optimal control with random disturbances. *Prikl. Mat. mekh.* 24, No. 1 (1960)
- 2 AOKI, M. Stochastic time-optimal systems. *Appl. Ind.*, No. 54 (May 1961)
- 3 FELDBAUM, A. A. Problems of the statistical theory of automatic optimization systems. *Automatic and Remote Control*, Vol. 2. 1960. London; Butterworths
- 4 NOVOSELTSEV, V. N. Optimal control in a second-order relay sampled-data system with random disturbances. *Automat. telemekh.* 22, No. 7 (1961)
- 5 FLORENTIN, J. J. Optimal control of continuous time, Markov and stochastic systems. *J. Electron. Contr.*, 1st Series. 10, No. 6 (1961)
- 6 NOVOSELTSEV, V. N. Time optimal control systems with random interference. *Automat. telemekh.* 23, No. 12 (1962)
- 7 FELDBAUM, A. A. *Computers in Automatic Control Systems*, 1959. Moscow; Fizmatgiz (in Russian)
- 8 SMITH, J. M. *Feedback Control Systems*, Pt. H. 1958. New York; McGraw-Hill

## DISCUSSION

W. M. WONHAM, *Research Institute for Advanced Studies, 7212 Bellona Avenue, Baltimore 12, M.D., U.S.A.*

If I understand the author correctly, he proposes to replace a stochastic control problem by a deterministic control problem which would be equivalent, in some approximate sense, to the original. The author suggests defining the 'equivalent' deterministic problem by using as a performance index 'the criterion of the minimum time of the transient process for the mathematical expectation'.

It is not clear to me precisely what the author means by this criterion, or by the corresponding value function  $T^*$ . There is a difficulty here which the author may have overlooked, namely, that the expected trajectories of a random process cannot, in general, be identified with the trajectories of some non-random dynamical system. In

other words, one does not have an 'equivalent' deterministic system to optimize.

To illustrate this, suppose that an admissible control law has been chosen, and that the resulting stochastic system generator is a homogeneous Markov process  $[x(t)]$  with the transition function<sup>1</sup>  $P(t, x, T)$ . The expected trajectory starting at a state  $x_0$  is by definition:

$$m(t, x_0) = \int x P(t, x_0) dx \quad t \geq 0$$

A non-random dynamical system, which is equivalent in the sense described, exists only if the function  $m$  has the semi-group property, namely

$$m(s+t, x_0) = m[t, m(s, x_0)], \quad s \geq 0, \quad t \geq 0 \quad (1^*)$$

Yet the Chapman-Kolmogorov equation<sup>1</sup> by no means implies the truth of (1\*), which is generally false, except when the stochastic system is linear. In fact, the author's actual procedure amounts to assuming that (1\*) is approximately valid when  $s$  and  $t$  are small: that is, of the order of the sampling interval  $\Delta$ . Yet it is precisely the validity of (1\*) for all  $s \geq 0$ ,  $t \geq 0$  which is necessary if dynamic programming arguments are to be relevant in this context.

Finally, I would welcome clarification of the following points of detail.

(1) If noise is present in the feedback channel, it is meaningless to seek a control law in the form of a function of the state. Yet the author seems to do just this, and indicates no filtering operation.

(2) In the dynamic equations (20), the plant input is completely decoupled from all but one of the state variables.

(3) What is the graph of the switch curve of eqn (27)?

(4) The result for the switching curve, eqn (30), seems difficult to interpret if  $\sum a_i p_i \leq 0$ . Yet the author nowhere states that the restriction  $\sum a_i p_i > 0$  may be necessary.

## Reference

<sup>1</sup> DYNKIN, E.-B. *Markovskie Protsessi*. 1963. Moscow; Fizmatgiz

V. N. NOVOSELTSEV, *in reply*

The difficulties which arise when the condition

$$m(s+t, x_0) = m[t, m(s, x_0)] \quad (1)$$

is not satisfied, can be avoided by examining the assembly average of the trajectory and also the assembly average of the trajectory with fixed initial conditions:

$$m(x, \tau) = \int_{\Omega(x)} x P(x_0, 0; x, \tau; u) dx \quad (2)$$

where  $P(x_0, 0; x, \tau; u)$  is the probability that at a moment of time  $\tau$  the system, under the influence of control  $u$ , will be in a state  $x$ , if at the initial moment  $t_0 = 0$  it was in the state  $x_0$ . The integral is taken for  $\Omega(x)$ , the whole domain of variation of  $x$ .

When in eqn (2)  $\tau$  rises continuously from 0 to  $\infty$  with a fixed  $u(t)$ ,  $0 \leq t \leq \infty$ , the geometrical place of the points  $m(x, \tau)$  in the phase space is the assembly average with initial condition  $x_0$ .

For some  $u(t)$  let trajectory (2) first reach the specified domain when  $\tau = 0$  (for some  $u(t)$  the quantity  $\theta$  can prove infinitely great). Then  $\theta^*$  is the minimal value of  $\theta$ , achieved when  $u(t) = u^0(t)$ :

$$\begin{aligned} \theta^* &= \min \theta[u(t)] \\ u(t) &\in \Omega[u(t)] \end{aligned} \quad (3)$$

As shown in the paper, the control algorithm

$$u^0(t) = u^0[x(t)]$$

ensuring optimality in regard to criterion  $T^*$ , i.e., ensuring the minimum  $T$  is also optimal in the sense of speed of response.

Although the proof of the latter assertion is based on relationship (9) in the paper, which is also used in deriving Bellman's equation, the dynamic programming method is not used in the paper.

If condition (1) is not satisfied, a system with prediction can be used as a convenient deterministic analogue of a stochastic system.

It is interesting to note that if the interference is a purely random process, the trajectory, although differing from the trajectory in a system without interference, proves to be non-random (this follows from the theory of large numbers<sup>1</sup>, and was noted by N.N. Krasovskii).

To the other questions I can reply as follows:

(1) If in the feedback channel (or the input channel) there is a filter, the variation of the interference distribution function must be taken into account. It is easy to show that the variation of the spectral properties of the interference does not change the algorithm, provided the current value of the interference is not being measured by the con-

troller. Since filtration of the error signal will also take place in such a system, a predictor must be connected ahead of the optimal controller<sup>2</sup>.

As before, the main element of such a more complex system remains the controller, the algorithm of which is determined in the paper.

(2) The paper considers a system in which the  $n$ th derivative of the output signal is constrained.

(3) The optimal switching line for a second-order system, with interference  $\eta_0$  and  $\eta_1$ , constrained in amplitude, present, is shown in Figure A. Here  $a$  is the optimal switching line in a system without interference,  $bc$  and  $de$  the optimal switching lines in the second and fourth quadrants with interference present,  $ABO$  the optimal trajectory without interference, and  $ACDEO$  the optimal trajectory with interference (for purely random noise the trajectory is the only one). Both trajectories contain two intervals.

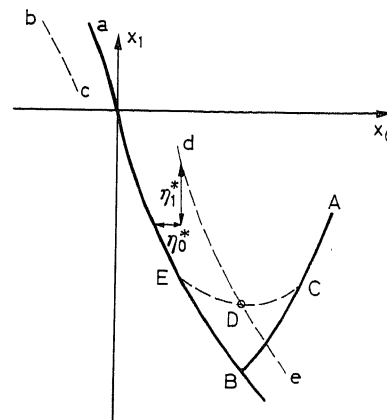


Figure A

(4) As a rule

$$\gamma = \sum_{i=1}^r a_i p_i > 0.$$

If, however,  $\gamma < 0$ , eqn (30) in the paper remains valid with  $\gamma$  replaced by  $|\gamma|$ , if the sign of the right-hand side of eqn (22) is changed. This is explained by the fact that the amount of information fed along the channel is determined by the modulus of quantity  $\gamma$ .

## References

<sup>1</sup> DUB, D. *Probability Processes*. 1956

<sup>2</sup> SMITH, O. J. M. *Automatic Control*. 1962. Fizmatgiz

R. E. KALMAN, *Research Institute for Advanced Studies, Centre for Control Theory, 7212 Bellona Avenue, Baltimore 12, M.D., U.S.A.*

Dr. Novoseltsev's ideas are quite interesting. Since his approach is a rather obvious one, a much deeper investigation would be required before the significance of his ideas can be assessed. There are three ways in which this could be done: (1) Rigorous mathematical methods; (2) Intuitive engineering arguments, together with a precise statement of unproved assumptions; and (3) Intuitive engineering arguments, together with a careful experimental verification of the assumptions. The paper does not meet any of these requirements.

The work of Bellman, upon which this paper is based, has been an extraordinarily stimulating influence. However, it is now generally recognized, especially since the contributions of Pontryagin and his collaborators, that the basic principles of dynamic programming are insufficient for practical applications and a deeper mathematical investigation is required in each case. (See the work of Bryson and Kelley in particular.) Especially difficult is the case of stochastic systems, where even very basic questions (existence and stability) are still largely unsolved. (See, however, the work of Gikhman and Krasovskii.)

The present opinion in the U.S.A. is that the theory of the effect of random disturbances, in non-linear dynamic systems, is presently at too primitive a level to warrant crude empirical applications of dynamic programming (as was done in this paper) to actual engineering problems.

In the absence of solid theoretical knowledge it would be dangerous to claim that practical problems can be attacked effectively by a crude simplification of the real problem. It is good to base engineering arguments on a simplified scientific theory, but simplicity in itself is no guarantee of validity. Our experience in the U.S.A. is that oversimplified approaches, as in this paper, do not really contribute to the solution of practical problems.

V. N. NOVOSELTSEV, *in reply*

Obviously the method of dynamic programming does not permit us at present to obtain a solution to the problem of synthesis of an optimal

algorithm (e.g., Reference 2 in the paper). It is therefore justifiable to seek other methods of solution, among which is the method examined in the paper. This is based on replacement of the ordinary optimum criterion by a simpler auxiliary criterion, and does not involve the use of dynamic programming, although based on the same initial relationships.

Thorough experimental verification and development of the results presuppose:

(a) Direct simulation of the optimal control algorithms according to the formulae obtained. Some results of such simulation are given in the paper.

(b) Static synthesis of optimal algorithms. Original work in this field is being done by R. I. Stakhovskii, using multi-channel automatic optimizers.

(c) Verification on prediction control systems, when the assembly average of the random trajectory in the future is predicted directly.

# DISCRETE SYSTEMS

## Quasi-invariant Hybrid Multi-parameter Control Loops

V. STREJC

### Summary

The article describes the general theory of the synthesis of hybrid, multi-parameter control systems acted upon by disturbances of an arbitrary form and with constant command input signals. The control loops contain controllers which operate in parallel and can be realized by means of an automatic digital computer and continuously acting controllers. A solution is presented according to which disturbances  $u_k$  effect only the controlled variables  $x_i$ ,  $i = k$ , and the degree of this effect can be limited in accordance with the selected criterion of the quality of control, or according to other suitable conditions of control. Control loops of this kind are called quasi-invariant control loops. The paper describes the application of the criterion of the finite number of control steps and of the criterion of the least square of deviations. The validity of the solution of the Wiener-Hopf integral equation is extended to control loops containing digital computers.

### Sommaire

Le rapport présente la théorie générale de la synthèse des systèmes de réglage hybrides à plusieurs variables, influencés par les perturbations d'une forme arbitraire et soumis à des grandeurs de commande constantes. Dans les systèmes de réglage sont incorporés en parallèle des régulateurs à action discontinue, réalisés par un calculateur numérique et des régulateurs. On a montré une solution d'après laquelle les perturbations  $u_k$  influencent seulement les grandeurs réglées  $x_i$ ,  $i = k$ , avec la possibilité de réduire le degré de cette influence d'après un critère choisi de la qualité de réglage ou d'après d'autres conditions appropriées. Ces systèmes de réglage sont désignés comme «systèmes quasi invariants». Dans ce travail on a appliqué le critère du nombre fini de pas de réglage et le critère du minimum de l'écart quadratique moyen et on a étendu la validité de la solution de l'équation intégrale de Wiener-Hopf à des systèmes de réglage contenant des calculateurs numériques.

### Zusammenfassung

Die Arbeit beschreibt eine allgemeine Theorie zur Synthese von gemischt analog-digital (hybrid) arbeitenden Mehrfachregelkreisen bei beliebigen Störgrößeneinwirkungen und konstanten Führungsgrößen. Die Regelkreise enthalten parallel arbeitende diskontinuierlich (Digitalrechner) und kontinuierlich wirkende Regler. Eine Lösung wird angegeben, bei der die Störgrößen  $u_k$  nur die Regelgrößen  $x_i$ ,  $i = k$  beeinflussen; der Grad dieser Beeinflussung läßt sich gemäß der gewählten Regelgüte oder nach anderen zweckmäßigen Bedingungen der Regelung beschränken.

Regelkreise dieser Art werden als „quasi-invariant“ bezeichnet. Der Aufsatz beschreibt die Anwendung der Methode der Regelung durch

eine endliche Anzahl von Schritten und das Kriterium der minimalen quadratischen Abweichung. Die Gültigkeit der Lösung der Wiener-Hopfschen Integralgleichung wird auf Regelkreise mit Digitalrechnern erweitert.

### Introduction

Previous papers by the author<sup>1-3</sup> contain the general theory of the synthesis of control systems and of the compensation of the effects of disturbances in hybrid, multi-parameter control loops, with due consideration of the conditions of autonomy, invariance and the finite number of control steps. A control loop is regarded as hybrid if the function of the controller is performed by a discrete filter (digital correcting member), the realization of which is assumed to be attainable by an automatic digital computer and a continuously-acting controller. In practice, hybrid control loops can be formed by the addition of an automatic computer to control loops containing continuously-acting controllers. This arrangement is made either in cases where it is necessary to improve the quality of control and to attain a higher stage of complex automation that would be difficult or too costly to realize by other means, or in newly designed control systems with the automatic computer as the main technical means of realizing automation and in which the simple, continuously-acting controllers are used as a stand-by for sustaining the operation of the control system in the case of an outage of the automatic computer.

In practical applications the case of a multi-parameter control system may occur frequently where the desired values of the controlled variables remain constant (their relative deviations

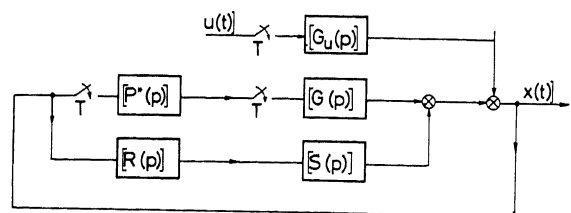


Figure 1

being zero), and the task of the control system is confined to the compensation of the effects of disturbances. If a control-system structure, according to Figure 1, is selected for a multi-parameter control loop of this kind, the conditions of invariance cannot be fulfilled. However, the existence of a solution will be presented according to which only the controlled variables  $x_i$ ,  $i = k$ , are influenced by disturbances  $u_k$  with the possibility of determining the limits of this influence, according to the selected criterion of the quality of control, or according to other suitable control conditions. Let control loops of this kind be designated as quasi-invariant control loops.

For a control loop according to Figure 1:

$$[K_u^*(z, 0)] = \{[1] + [\Omega^*(z, 0)][P^*(z, 0)]\}^{-1} [\Omega_u^*(z, 0)] \quad (1)$$

$$[K_u^*(z, \varepsilon)] = [\Omega_u^*(z, \varepsilon)] - [\Omega^*(z, \varepsilon)][P^*(z, 0)][K_u^*(z, 0)] \quad (2)$$

where

$$[\Omega(p)] = \{[1] + [S(p)][R(p)]\}^{-1} [G(p)]$$

$$[\Omega_u(p)] = \{[1] + [S(p)][R(p)]\}^{-1} [G_u(p)]$$

In the compensation of disturbance effects  $[K_u^*(z, \varepsilon)]$  and  $[K_u^*(z, 0)]$  are the matrices of the transfer functions of closed control loops with the elements of the matrices expressed as discrete Laplace transforms ( $Z$  transforms). In eqn (2)  $\varepsilon$  stands for the relative value of the independent time variable which during one interval of sampling attains the value of  $\varepsilon < 0 \div 1 >$ . The sampling interval  $T$  is constant, and let the sampling be synchronous at all points of the control loop.

$[u(i)]$  is the  $(\xi; 1)$  type column matrix of the disturbances

$[x(i)]$  is the  $(\nu; 1)$  type column matrix of the controlled variables

$[S(p)]$  is the  $(\nu; \mu)$  type rectangular matrix,  $\mu \geq \nu$  of the transfer functions of the controlled system containing a servomotor and a final control element

$[G(p)]$  is the  $(\nu; \mu)$  type rectangular matrix,  $\mu \geq \nu$ , of the transfer functions of a controlled system containing a servomotor, final control element and a holding member

$[G_u(p)]$  is the  $(\nu; \xi)$  type rectangular matrix,  $\xi \leq \nu$ , of the transfer functions of the controlled system containing a holding member

$[P^*(p)]$  is the  $(\mu; \nu)$  type rectangular matrix,  $\mu \geq \nu$ , of the transfer functions of digital correcting members

$[R(p)]$  is the  $(\mu; \nu)$  type rectangular matrix,  $\mu \geq \nu$ , of the transfer functions of continuously-acting controllers

### Conditions of Stability

As it is desirable to express the quality of control by the requirements upon the transfer functions in matrix  $[K_u^*(z, 0)]$ , the matrix  $[P^*(z, 0)]$  is a function of matrix  $[K_u^*(z, 0)]$ . It can be calculated from eqn (1) that

$$[P^*(z, 0)] = [\Omega^*(z, 0)]^{-1} \{[\Omega_u^*(z, 0)] - [K_u^*(z, 0)]\} [K_u^*(z, 0)]^{-1} \quad (3)$$

By substituting relation (3) for  $[P^*(z, 0)]$  into eqn (2)

$$[K_u^*(z, \varepsilon)] = [\Omega_u^*(z, \varepsilon)] - \frac{\Delta_{\Omega A}(z, 0)}{\Delta_{\Omega B}(z, 0)} [\Omega^*(z, \varepsilon)] [\omega^*(z, 0)] \{[\Omega_u^*(z, 0)] - [K_u^*(z, 0)]\} \quad (4)$$

where

$$\frac{\Delta_{\Omega A}(z, 0)}{\Delta_{\Omega B}(z, 0)} [\omega^*(z, 0)] = [\Omega^*(z, 0)]^{-1} \quad (5)$$

As the continuously-acting controllers, the transfer functions of which have the matrix  $[R(p)]$ , are determined by the condition of all loops of the control system remaining stable in the case of a computer outage, it may be stated that the elements of matrix  $[\Omega_u^*(z, \varepsilon)]$  will always be stable.

On the other hand, the elements of the second term on the right-hand side of eqn (4) can be unstable if the polynomial  $\Delta_{\Omega B}(z, 0)$ , which is the numerator of the determinant  $\dagger \Delta_{\Omega}(z, 0)$  of matrix  $[\Omega^*(z, 0)]$ , has its zero outside the zone of stability. The unstable zeros of polynomial  $\Delta_{\Omega B}(z, 0)$  must be assumed, however, to be compensated by the numerator of the elements of matrix  $\{[\Omega_u^*(z, 0)] - [K_u^*(z, 0)]\}$  which, as can be seen from eqn (3), is a cofactor of the matrix  $[P^*(z, 0)]$ . In accordance with the assumptions stated previously, the elements of matrices  $[\Omega^*(z, \varepsilon)]$  and  $[\omega^*(z, 0)]$  in eqn (4) are stable, while the stability of the elements of matrix  $[K_u^*(z, 0)]$  must be presupposed. On the basis of the above findings, it is possible to state the condition of the stability of a hybrid, multi-parameter control system for the compensation of disturbance effects, as follows:

$$[\Omega_u^*(z, 0)] - [K_u^*(z, 0)] = \Delta_{\Omega B}^-(z, 0) [D_u^*(z, 0)] \quad (6)$$

where  $[D_u^*(z, 0)]$  is the matrix of auxiliary functions that must be determined in more detail, while  $\Delta_{\Omega B}^-(z, 0)$  follows from the equation

$$\Delta_{\Omega B}(z, 0) = \Delta_{\Omega B}^+(z, 0) \Delta_{\Omega B}^-(z, 0) \quad (7)$$

where  $\Delta_{\Omega B}^+(z, 0)$  signifies the product of the stable, and  $\Delta_{\Omega B}^-(z, 0)$  the product of the unstable root factors of the numerator of the determinant of matrix  $[\Omega^*(z, 0)]$ .

Introduce

$$[K_u^*(z, 0)] = [\bar{Q}_u^*(z, 0)] [(1 - z^{-1})^m C^*(z, 0)] \quad (8)$$

where  $[\bar{Q}_u^*(z, 0)]$  is the matrix of auxiliary functions, and  $[(1 - z^{-1})^m C^*(z, 0)]$  is a diagonal matrix of the  $(\xi; \xi)$  type, the elements of which should be polynomials independent of the properties of control loop members. Let these elements be the denominators of the  $Z$  transforms of the general form of the disturbances

$$[U^*(z, 0)] = \left[ \frac{F^*(z, 0)}{(1 - z^{-1})^m C^*(z, 0)} \right] \quad (9)$$

Now, eqn (6) can be rewritten in the form

$$[\Omega_u^*(z, 0)] - [\bar{Q}_u^*(z, 0)] [(1 - z^{-1})^m C^*(z, 0)] = \Delta_{\Omega B}^-(z, 0) [D_u^*(z, 0)] \quad (10)$$

After substituting relations (8) and (10) into eqn (3), matrix  $[P^*(z, 0)]$  expressed by this equation will acquire the form

$$[P^*(z, 0)] = \frac{\Delta_{\Omega A}(z, 0)}{\Delta_{\Omega B}^+(z, 0)} [\omega^*(z, 0)] [D_u^*(z, 0)] \{[\bar{Q}_u^*(z, 0)] [(1 - z^{-1})^m C^*(z, 0)]\}^{-1} \quad (11)$$

$\dagger$  The 'determinant' of the  $(m; n)$  type rectangular matrix  $A$ , with  $m < n$ , is to be considered as being identical with the determinant of matrix  $A^T A$  where  $A^T$  is the matrix transposed towards matrix  $A$ .

Now, let the real functions in eqn (2) be marked with the subscript  $s$ , and the imaginary functions in eqn (11) with the subscript  $p$ . After substituting relation (11) into eqn (2), it follows

$$[K_{us}^*(z, \varepsilon)] = [\Omega_{us}^*(z, \varepsilon)] - \frac{\Delta_{\Omega A}(z, 0)}{\Delta_{\Omega B}(z, 0)} [\Omega_s^*(z, \varepsilon)] [\omega_p^*(z, 0)] [D_{up}^*(z, 0)] \quad (12)$$

provided that

$$[\bar{Q}_{up}^*(z, 0)]^{-1} [\bar{Q}_{us}^*(z, 0)] = [1] \quad (13)$$

Assumption (13) can be fulfilled only if the zeros and the poles of the determinant of matrix  $[Q_u^*(z, 0)]$  are inside the zone of stability.

The following holds for the elements of matrices in eqn (10):

$$\begin{aligned} \Omega_{u, ik}^*(z, 0) - \bar{Q}_{u, ik}^*(z, 0) (1 - z^{-1})_{kk}^m C_{kk}^*(z, 0) \\ = \Delta_{\Omega B}^-(z, 0) D_{u, ik}^*(z, 0) \end{aligned} \quad (14)$$

Introduce

$$\left. \begin{aligned} K_{u, ik}^*(z, 0) &= \frac{K_{uB, ik}^*(z, 0)}{K_{uA, ik}^*(z, 0)} & \bar{Q}_{u, ik}^*(z, 0) &= \frac{\bar{Q}_{uB, ik}^*(z, 0)}{\bar{Q}_{uA, ik}^*(z, 0)} \\ \Omega_{u, ik}^*(z, 0) &= \frac{M_{uB, ik}^*(z, 0)}{M_{uA, ik}^*(z, 0)} & D_{u, ik}^*(z, 0) &= \frac{D_{uB, ik}^*(z, 0)}{D_{uA, ik}^*(z, 0)} \end{aligned} \right\} \quad (15)$$

where fractions (15) represent the ratios of polynomials in  $z^{-1}$  with a finite number of terms. By using relations (15), eqn (14) can be rewritten in the form

$$\begin{aligned} \frac{M_{uB, ik}^*(z, 0)}{M_{uA, ik}^*(z, 0)} - \frac{\bar{Q}_{uB, ik}^*(z, 0)}{\bar{Q}_{uA, ik}^*(z, 0)} (1 - z^{-1})_{kk}^m C_{kk}^*(z, 0) \\ = \Delta_{\Omega B}^-(z, 0) \frac{D_{uB, ik}^*(z, 0)}{D_{uA, ik}^*(z, 0)} \end{aligned} \quad (16)$$

Similarly, the following holds for the elements of the matrices in eqn (8)

$$\frac{K_{uB, ik}^*(z, 0)}{K_{uA, ik}^*(z, 0)} = \frac{\bar{Q}_{uB, ik}^*(z, 0)}{\bar{Q}_{uA, ik}^*(z, 0)} (1 - z^{-1})_{kk}^m C_{kk}^*(z, 0) \quad (17)$$

In the case of

$$D_{uA, ik}^*(z, 0) = M_{uA, ik}^*(z, 0) \bar{Q}_{uA, ik}^*(z, 0) \quad (18)$$

eqn (16) will acquire the form

$$\begin{aligned} \bar{Q}_{uA, ik}^*(z, 0) M_{uB, ik}^*(z, 0) \\ - \bar{Q}_{uB, ik}^*(z, 0) M_{uA, ik}^*(z, 0) (1 - z^{-1})_{kk}^m C_{kk}^*(z, 0) \\ = \Delta_{\Omega B}^-(z, 0) D_{uB, ik}^*(z, 0) \end{aligned} \quad (19)$$

Denote

$$\left. \begin{aligned} \Delta_{\Omega B}^-(z, 0) &= 1 + \sum_{v=1}^{L_1} b_v z^{-v} & \bar{Q}_{uA, ik}^*(z, 0) &= 1 + \sum_{v=1}^{Q_A} p_v z^{-v} \\ C_{kk}^*(z, 0) &= 1 + \sum_{v=1}^C c_v z^{-v} & \bar{Q}_{uB, ik}^*(z, 0) &= \sum_{v=1}^{Q_B} q_v z^{-v} \\ M_{uA, ik}^*(z, 0) &= 1 + \sum_{v=1}^{M_A} \alpha_v z^{-v} & D_{uB, ik}^*(z, 0) &= \sum_{v=1}^{D_B} d_v z^{-v} \\ M_{uB, ik}^*(z, 0) &= \sum_{v=1}^{M_B} \beta_v z^{-v} \end{aligned} \right\} \quad (20)$$

Let the degree of polynomial  $\bar{Q}_{uB, ik}^*(z, 0)$  be assumed as

$$Q_B = Q + N \quad (21)$$

where  $Q$  is the lowest possible degree of the polynomial  $\bar{Q}_{uB, ik}^*(z, 0)$  that follows from eqn (19), and  $N$  the number of degrees of freedom.

Assuming that

$$Q_A + M_B \leq Q_B + M_A + m + C \quad (22)$$

the degree of the resultant polynomial on the left-hand side of eqn (19) will be

$$Q_B + M_A + m + C = L_1 + D + N \quad (23)$$

From eqn (23) follows the degree of polynomial  $\bar{Q}_{uB, ik}^*(z, 0)$

$$Q_B = L_1 + N = Q + N$$

and the degree of polynomial  $D_{uB, ik}^*(z, 0)$

$$D = M_A + m + C \quad (24)$$

$$D_B = D + N \quad (25)$$

By comparing the coefficients of the equal powers  $z^{-1}$  in the resultant polynomials on both sides of eqn (19), the system of  $Q_B + D$  linear algebraic equations is obtained where

$$Q_B + D = L_1 + D + N \quad (26)$$

To this system of equations it is necessary to add further  $N + Q_A$  equations of conditions that follow from the selected conditions of control. The system of  $Q_B + D_B + Q_A$  equations obtained in this way determines the coefficients of polynomials  $\bar{Q}_{uB, ik}^*(z, 0)$ ,  $\bar{Q}_{uA, ik}^*(z, 0)$  and  $D_{uB, ik}^*(z, 0)$  of the auxiliary functions, provided that the determinant of the equation system does not equal zero. The number of such coefficients is

$$Q_A + Q_B + D_B = L_1 + D + 2N + Q_1 \quad (27)$$

A more detailed analysis would prove that  $\beta_1 = q_1$  and  $d_1 = 0$  holds generally, and consequently the number of conditions necessary for the determination of the coefficients of auxiliary functions may be reduced by two.

The solution is somewhat simplified if it can be stated that

$$K_{uA, ik}^*(z, 0) = \bar{Q}_{uA, ik}^*(z, 0) = M_{uA, ik}^*(z, 0) \quad (28)$$

It follows

$$D_{uA, ik}^*(z, 0) = M_{uA, ik}^*(z, 0) \quad (29)$$

and eqn (16) will assume the form

$$\begin{aligned} M_{uB, ik}^*(z, 0) - \bar{Q}_{uB, ik}^*(z, 0) (1 - z^{-1})_{kk}^m C_{kk}^*(z, 0) \\ = \Delta_{\Omega B}^-(z, 0) D_{uB, ik}^*(z, 0) \end{aligned} \quad (30)$$

The above simplification does not allow the inclusion, in the characteristic equation of transfer functions  $K_{u, ik}^*(z, 0)$ , of additional requirements above those asserted in the characteristic equation of the terms  $\Omega_{u, ik}^*(z, 0)$ .

After the determination of all elements of matrix  $[\bar{Q}_u^*(z, 0)]$  it is necessary to check the zeros in the numerator of the determinant of this matrix.

Now, the conditions of stability can be summarized as:

**Theorem 1**—In the defined hybrid control loop where  $\Delta_{\Omega B}^-(z, 0)$  is the product of the unstable root factors of the numerator of

the determinant of matrix  $[\Omega^*(z, 0)]$ , with the poles of this determinant lying within the stable zone of plane  $z$ , the transfer functions of the control loops, i.e. the elements of matrix  $[K_u^*(z, \varepsilon)]$ , are stable, provided that: (a) the poles and zeros of the determinant of matrix  $[\bar{Q}_u^*(z, 0)]$  lie within the stable zone of plane  $z$ , (b) the matrix  $[\bar{Q}_u^*(z, 0)]$  is in accordance with the equation of conditions (10) and none of its elements is equal to zero, and (c) the poles of the elements of matrix  $[D_u^*(z, 0)]$  in eqn (10) also lie in the stable zone of plane  $z$ . These conditions are necessary and sufficient.

### The Conditions of Zero Offset

Provided that the conditions of stability are fulfilled, it is possible to state the condition of zero offset according to the theorem of finite values by the following equation:

$$\lim_{z \rightarrow 1} K_{u, ik}^*(z, 0) = 0 \quad (31)$$

The above condition can be fulfilled if the value of  $m$  in the general relation (8) is at least  $m = 1$ . In other words, the product of the numerator root factors of transfer functions  $K_{u, ik}^*(z, 0)$  must necessarily contain the factor  $(1 - z^{-1})$ .

### Quasi-invariant Control Loops

Provided that all zeros of the determinant of matrix  $[\Omega^*(z, 0)]$  lie within the stable zone of plane  $z$ , the following substitution can be made in eqn (14):

$$\Delta_{\Omega B}^-(z, 0) = 1 \quad (32)$$

If the selectable functions are stated as

$$\bar{Q}_{u, ik}^*(z, 0) = 0 \quad \text{for } i \neq k \quad (33)$$

it follows

$$D_{u, ik}^*(z, 0) = \Omega_{u, ik}^*(z, 0) \quad \text{for } i \neq k \quad (34)$$

and the remaining functions  $D_{u, ik}^*(z, 0)$ ,  $i = k$ , can be determined by the same method as shown earlier in this paper. In this case the matrix  $[\bar{Q}_u^*(z, 0)]$  will be a diagonal matrix and consequently, with regard to eqn (8),  $[K_u^*(z, 0)]$  will also be a diagonal matrix. This solution permits a situation to be reached where disturbances  $U_k^*(z, 0)$ , ( $k = 1, 2, \dots, \xi$ ), (where  $\xi \geq \nu$  and  $\nu$  is the number of controlled variables) will influence only the controlled variables  $X_i^*(z, \varepsilon)$ , for which  $i = k$ , and will have no influence upon the controlled variables  $X_i^*(z, \varepsilon)$ , for which  $i \neq k$ . If  $\xi < \nu$ , the effect of disturbances  $U_k^*(z, 0)$  will be confined to the controlled variables  $X_i^*(z, \varepsilon)$ , for which  $i = k$  and  $i = 1, 2, \dots, \xi$  and with no effect upon the controlled variables  $X_i^*(z, \varepsilon)$ , for which  $i \neq k$  and also those for which  $i = k$  but  $i = \xi + 1, \xi + 2, \dots, \nu$ .

Due to this solution the transfer functions in diagonal matrix  $[K_u^*(z, 0)]$  can have an arbitrary number of degrees of freedom that can be utilized for the fulfilment of further conditions of control, or for the compliance with a suitable criterion of the quality of control. In this way it is possible to reach a solution at which the effect of disturbances, that cannot be eliminated by the introduction of condition (33), is kept within admissible limits.

In principle, this method of the compensation of disturbance effects can also be applied to continuously-acting control systems. However, up to the time of writing this paper, this possibility has not been mentioned in any technical literature accessible to the author.

### Finite Number of Control Steps

In the compensation of disturbance effects a multi-parameter control loop complies with the requirement of a finite number of control steps, if the same requirement is complied with by all components of output signals  $X_i^*(z, \varepsilon)$  of the controlled system. In this case

$$X_{ik}^*(z, \varepsilon) = K_{u, ik}^*(z, \varepsilon) U_k^*(z, 0) \quad (35)$$

The finite number of control steps is understood as the number of sampling intervals, at the attainment of which the offset is permanently zero or constant at any one instant of sampling. In the intervals between the instants of sampling this condition need not be fulfilled. If it is possible to express the  $Z$  transforms of the general forms of the disturbances by eqn (9), and the matrix of transfer functions  $[K_u^*(z, 0)]$  by eqn (8), it follows

$$X^*(z, 0) = [\bar{Q}_u^*(z, 0)] F^*(z, 0) \quad (36)$$

It follows from eqn (36) that the requirement of the finite number of control steps can be complied with only if the elements of matrix  $[\bar{Q}_u^*(z, 0)]$  are polynomials having a finite number of terms.

In this case it is necessary to substitute in eqns (18) and (19)

$$\bar{Q}_{uA, ik}^*(z, 0) = 1 \quad (37)$$

Then for individual components

$$X_{ik}^*(z, 0) = \bar{Q}_{u, ik}^*(z, 0) F_k^*(z, 0) \quad (38)$$

and the degree of polynomial  $X_{ik}^*(z, 0)$  is

$$X = L_1 + N + F \quad (39)$$

The transform of the controlled variable is

$$X_i^*(z, 0) = \sum_{k=1}^{\xi} X_{ik}^*(z, 0) \quad (40)$$

with the degree of polynomial  $X_i^*(z, 0)$  being

$$X_i = L_1 + (N + F)_i \quad (41)$$

where  $(N + F)_i$  is the highest value of the sum  $N + F$  in the polynomials  $X_{ik}^*(z, 0)$ , ( $k = 1, 2, \dots, \xi$ ).

The number of the control steps is thus

$$n_{ki} = L_1 + (N + F)_i + 1 \quad (42)$$

The highest value of  $n_{ki}$  ( $i = 1, 2, \dots, \nu$ ), is regarded to be the finite number of the control steps of the whole multi-parameter system.

If disturbances  $u_k(t)$  can be regarded as the linear combination of the function  $t^{m-1}/(m-1)!$ , it follows  $F = m - 1$  and the number of control steps is

$$n_{ki} = L_1 + (N + m)_i \quad (43)$$

It can equally be proved that it is also possible to obtain a zero deviation of the controlled variables  $X_i^*(z, \varepsilon)$  for  $\varepsilon < 0 < 1$  beginning with the instant  $n = n_{ki}$  provided that: disturbance  $u_k(t)$  varied from instant  $n = 0$  according to the function  $u_k(t) = t^{m-1}/(m-1)!$ , the elements of matrices  $[G(p)]$  and  $[G_u(p)]$  have at least one  $m$ -fold zero pole (in hybrid loops the elements of matrices  $[G(p)]$  and  $[G_u(p)]$  must have a holding member at least of the order  $(m - 1)$ ), in the equation of conditions (6) and

in the equations derived from it  $\Delta_{\Omega B}(z, 0)$  is substituted for  $\Delta_{\Omega B}(z, 0)$  and the auxiliary functions in matrices  $[D_u^*(z, 0)]$  and  $[Q_u^*(z, 0)]$  are only polynomials in  $z^{-1}$ .

Then it follows from eqn (16)

$$\begin{aligned} M_{uB, ik}^*(z, 0) \\ - \bar{Q}_{uB, ik}^*(z, 0)(1 - z^{-1})^{m-1} C_{kk}^*(z, 0) M_{uA, ik}^*(z, 0) \\ = \Delta_{\Omega B}(z, 0) D_{uB, ik}^*(z, 0) M_{uA, ik}^*(z, 0) \end{aligned} \quad (44)$$

The degree of eqn (44) is

$$Q_B + m - 1 + C + M_A = l_1 + D + N + M_A \quad (45)$$

$$D_B = D + N \quad (46)$$

$$D = m - 1 + C + M_A \quad (47)$$

$$Q_B = l_1 + N + M_A \quad (48)$$

The number of control steps is then

$$n_{ki} = (Q_B + F)_i \quad (49)$$

$$n_{kl} = l_1 + (M_A + N + C + m)_i - 1 \quad (50)$$

where  $l_1$  is the degree of the polynomial  $\Delta_{\Omega B}(z, 0)$ .

Let it be noted further that in eqn (44)

$$D_{uB, ik}^*(z, 0) = \sum_{v=0}^D d_v z^{-v} \quad (51)$$

differently from the polynomial in eqn (20). Owing to  $d_0 \neq 0$  it has been possible to reduce exponent  $m$  in eqn (44). It should also be mentioned that in this case the number of control steps cannot generally be lowered by the value of  $M_A$  by setting

$$D_{uB, ik}^*(z, 0) = \frac{D_{uB, ik}^*(z, 0)}{M_{uA, ik}^*(z, 0)}$$

because the output signal of the digital correction members is

$$E_2^*(z, 0) = -\Delta_{\Omega A}(z, 0) [\omega^*(z, 0)] [D_u^*(z, 0)] U^*(z, 0) \quad (52)$$

and it cannot be assumed that in a general case  $M_{uA, ik}^*(z, 0)$  is contained in  $\Delta_{\Omega A}(z, 0)$ . The denominators of the elements of matrix  $[\omega^*(z, 0)]$  are contained in  $\Delta_{\Omega A}(z, 0)$ .

### The Optimum Compensation of Disturbance Effects in Wiener's Sense

A method has been shown for limiting the effect of disturbances in quasi-variant control loops by the criterion of the finite number of control steps being considered as the criterion of the quality of control. Another method of solution will be shown where the least square of the deviations of the controlled variables is taken as the criterion of the quality of control. Let the problem be stated by the application of the conventional diagram shown in Figure 2 with the following meaning of denotations:

$[u(t)]$  = stationary random disturbances

$[m(t)]$  = parasitic noise

$[K_u^*(z, \varepsilon)]$  = the  $(\nu; \xi)$  type rectangular matrix of the transfer functions of the control system that are to be determined

$[I^*(z, \varepsilon)]$  = the  $(\nu; \xi)$  type rectangular matrix of the ideal transfer functions of the control system

$[\Delta(t)]$  = the deviations of the controlled variables  $x[(t)]$  from the ideal output signals  $[y(t)]$ .

$$\Delta_i(n, \varepsilon) = x_i(n, \varepsilon) - y_i(n, \varepsilon) \quad (53)$$

For the sake of brevity the analysis that follows deals only with the case of non-correlated input signals. By using the results published in an earlier paper by the author, the transfer functions sought for, i.e. the elements of matrix  $K_u^*(z, \varepsilon)$ , can be determined by the solution of equation

$$K_{u, ik}^*(j\bar{\omega}, \varepsilon) {}^1S_{kk}^*(\bar{\omega}, 0) - \Gamma_{ik}^*(j\bar{\omega}, \varepsilon) {}^2S_{kk}^*(\bar{\omega}, 0) = 0 \quad (54)$$

where  $K_{u, ik}^*(j\bar{\omega}, \varepsilon)$  and  $\Gamma_{ik}^*(j\bar{\omega}, \varepsilon)$  are the  $z$  transforms of the above-mentioned transfer functions,  $z = e^{j\bar{\omega}}$ ,  $\bar{\omega} = \omega T$  while  ${}^1S_{kk}^*(\bar{\omega}, 0)$  and  ${}^2S_{kk}^*(\bar{\omega}, 0)$  are discrete forms of the performance spectral densities:

$${}^1S_{kk}^*(\bar{\omega}, 0) = S_{u_k u_k}^*(\bar{\omega}, 0) + S_{m_k m_k}^*(\bar{\omega}, 0) \quad (55)$$

$${}^2S_{kk}^*(\bar{\omega}, 0) = S_{u_k u_k}^*(\bar{\omega}, 0) \quad (56)$$

Eqn (54) is representing the discrete Laplace transform of the Wiener-Hopf integral equation, the solution of which

$$K_{u, ik}^*(j\bar{\omega}, \varepsilon) = \frac{\left[ \frac{\Gamma_{ik}^*(j\bar{\omega}, \varepsilon) {}^2S_{kk}^*(\bar{\omega}, 0)}{{}^1S_{kk}^-(\bar{\omega}, 0)} \right]_+}{{}^1S_{kk}^+(\bar{\omega}, 0)} \quad (57)$$

by the known method fulfils the condition

$$k_{u, ik}(n, \varepsilon) = 0 \quad \text{for } n < 0 \quad (58)$$

where  $k_{u, ik}(n, \varepsilon)$  is the original of the transform  $K_{u, ik}^*(j\bar{\omega}, \varepsilon)$ . In eqn (57)

$${}^1S_{kk}^+(\bar{\omega}, 0) {}^1S_{kk}^-(\bar{\omega}, 0) = {}^1S_{kk}^*(\bar{\omega}, 0) \quad (59)$$

where all the poles of  ${}^1S_{kk}^+(\bar{\omega}, 0)$  are inside, and all the poles of  ${}^1S_{kk}^-(\bar{\omega}, 0)$  are outside the zone of stability of plane  $z$ . The  $+$  sign in the place of a subscript of the brackets in the numerator of eqn (57) signifies that the function in the brackets has all its poles inside the stability zone of plane  $z$ .

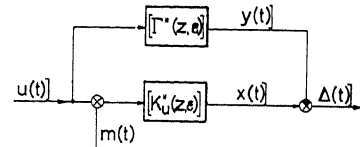


Figure 2

For  $\varepsilon = 0$  the elements of matrix  $[K_u^*(z, 0)]$  can be determined from eqn (57), and subsequently the matrix of the transfer functions of the digital correcting members is determined from eqn (3).

Equation (2) represents the unequivocal relationship that exists between transfer functions  $K_{u, ik}^*(z, 0)$  and  $K_{u, ik}^*(z, \varepsilon)$ , with the former necessarily fulfilling eqn (6) and also the conditions of stability attached to this equation. In order to ensure the stability of transfer functions  $K_{u, ij}^*(z, \varepsilon)$ , it is necessary



that the conditions of stability related to eqn (6) are also fulfilled by transfer functions  $K_{w,ij}^*(z,0)$ . This requirement can be complied with in various ways, one of which is stated below.

If, on the other hand, transfer functions  $K_{w,ij}^*(z,0)$  were not fulfilling the stability conditions related to eqn (6), the required course of the controlling actions would be ensured only at the instants of sampling by the digital correcting members calculated from eqn (3) and with the aid of transfer functions  $K_{u,ik}^*(z,0)$  determined by eqn (57). However, during the periods between the sampling instants, the course in time of the controlled variables cannot be guaranteed, and it may even be labile.

The new concepts can be summarized in the following theorem.

**Theorem 2**—For the determination of the digital correcting members in Wiener's sense, i.e. in a control loop containing a continuously-acting controlling system and exposed to the effects of stationary random disturbances, the mean square of the deviations of the real output signals from the ideal output signals should attain its minimum value, the command transfer functions  $K_{u,ik}^*(z,0)$  must fulfil the conditions of stability issuing from the solution of the Wiener-Hopf integral equation, and, in order to ensure the stability of transfer functions  $K_{u,ik}^*(z,\varepsilon)$ , the transfer functions  $K_{u,ik}^*(z,0)$  of this solution must comply with the conditions of stability pertaining to eqn (6).

These conditions are necessary and sufficient. This is the

fundamental difference between the described control loops and control loops containing only discretely-acting or only continuously-acting members.

The conditions stipulated in Theorem 2 can be fulfilled, if the root factors in the numerators and denominators of transfer functions  $K_{u,ik}^*(z,0)$  derived from the relations  ${}^1S_{kk}^*(\bar{\omega},0)$  and  ${}^2S_{kk}^*(\bar{\omega},0)$  are introduced as a condition into auxiliary functions  $\bar{Q}_{u,ik}^*(z,0)$  and  $D_{u,ik}^*(z,0)$  calculated from the equation of conditions (6). From this point of view, the solution according to the least square of deviations in Wiener's sense represents only the utilization of a possible application of the required criterion of quality of control within the determinative synthesis theory, and the possibility of extending the auxiliary functions  $\bar{Q}_{u,ik}^*(z,0)$  and  $D_{u,ik}^*(z,0)$  by the required number of degrees of freedom.

## References

- 1 STREJC, V. Ensuring reliability in complex automation by automatic digital computers. *Automatizace* 5 (1962) 123-125
- 2 STREJC, V., and RŮŽIČKA, J. Pt 1: *Izv. Akad. Nauk SSSR, OTN. Energetika i Avtomatika* 5 (1961. 57-71. Pt 2: *Izv. Akad. Nauk SSSR, OTN. Energetika i Avtomatika* (1963). In the press
- 3 STREJC, V. The theory of the synthesis of multi-parameter control systems containing automatic computers and acted upon by random signals. *Sbornik 9. Stroje na zpracovani informaci*. 1963. Prague; NCSAV. In the press

# Optimum Control of Discrete-time Dynamic Processes

B. FRIEDLAND

## Summary

The first part of the paper deals with the general problem of forcing a non-linear process to follow a specified (discrete-time) trajectory. The performance is measured by a functional of the vector error and input sequences and the inputs are assumed to be constrained. By the use of standard methods of calculus the optimum controller is shown to require the solution of a system of difference equations of twice the order of the process to be controlled with boundary conditions at the terminal instant dependent on the performance measure. A simple geometric interpretation of the problem is given.

The second part of the paper is concerned with a more specific class of problem in which the process is linear, and the performance objective is one of the following:

- (1) To force the process from an arbitrary initial state as close as possible to a specified terminal state in a fixed number of steps.
- (2) To force the process from the initial state to a specified terminal state in a fixed number of steps with minimum 'effort'.
- (3) To force the process from the initial state to a specified terminal state in a minimum number of steps.

A special-purpose computer which performs the required computation in real time is proposed as the on-line optimum controller.

## Sommaire

La première partie de la communication traite du problème général consistant à obliger un processus non-linéaire à suivre une trajectoire spécifiée en temps discret. La performance est mesurée à l'aide d'une fonctionnelle de l'erreur de vecteur et des séquences d'entrée et les signaux d'entrée sont supposés limités. Il est montré, à l'aide de méthodes de calcul habituelles, que le régulateur optimal exige la résolution d'un système d'équations aux différences finies dont l'ordre est doublé de celui du processus devant être réglé, avec des conditions aux limites à l'instant final dépendant de la mesure de la performance. Une interprétation géométrique simple du problème est donnée.

La seconde partie de la communication se rapporte à une catégorie plus spécifique de problèmes, dans laquelle le processus est linéaire et l'objectif de la performance est l'un des suivants:

- (1) Amener le processus d'un état initial arbitraire au voisinage le plus proche possible d'un état final déterminé en un nombre d'étapes donné.
- (2) Amener le processus de l'état initial à un état final déterminé en un nombre d'étapes donné avec l'«effort» minimal.
- (3) Amener le processus de l'état initial à un état final déterminé en un nombre d'étapes minimal. Un ordinateur à usage spécial qui effectue le calcul nécessaire en temps réel est proposé en tant que régulateur optimal en ligne.

## Zusammenfassung

Der erste Teil der Arbeit befaßt sich mit dem allgemeinen Problem, einen nichtlinearen Prozeß zu zwingen, einer festgelegten Trajektorie (für diskrete Zeitpunkte) zu folgen. Der Verlauf wird auf Grund des Verhaltens des Vektors für den Fehler und der Eingangsfolgen gemessen; dabei ist angenommen, daß die Eingangsgrößen beschränkt sind. Die Anwendung bekannter Berechnungsmethoden für optimale Regler zeigt, daß die Lösung eines Systems von Differenzgleichungen von der zweifachen Ordnung des zu regelnden Prozesses erforderlich

ist; die Grenzbedingungen im Endzustand sind dabei von dem gemessenen Verlauf abhängig. Eine einfache geometrische Deutung des Problems wird gegeben.

Der zweite Teil der Arbeit beschäftigt sich mehr mit einer speziellen Klasse von Problemen, in denen für einen linearen Prozeß einer der folgenden Gesichtspunkte zugrunde liegt:

Der Prozeß wird gezwungen:

1. von einem willkürlichen Anfangszustand aus so genau wie möglich einen festgelegten Endzustand in einer bestimmten Anzahl von Schritten zu erreichen,
2. von dem Anfangszustand zu einem festgelegten Endzustand in einer bestimmten Anzahl von Schritten mit minimalem Aufwand zu kommen und
3. von dem Anfangszustand zu einem festgelegten Endzustand in einer minimalen Anzahl von Schritten zu gelangen.

Ein spezieller Rechner, der die notwendigen Berechnungen in Echtzeit ausführt, wird als ein direkt arbeitender (on-line) optimaler Regler vorgeschlagen.

## Introduction

Traditionally, feedback control systems have been designed by the use of techniques in which the configuration of the controller is assumed and cut-and-try procedures are used to determine the settings of one or more adjustable parameters, such as gains, lag or lead times, within the controller. These techniques have proved to be inadequate for many problems in which the process to be controlled is non-linear, in which the actuating forces are constrained, and where high performance is essential. Consequently there has been an increasing interest in the development of a systematic theory of optimum control which does not suffer from these shortcomings. Numerous investigations carried out during the past decade with the aid of such mathematical tools as variational calculus (including its modification by Pontryagin and his students), functional analysis, and dynamic programming have uncovered results of considerable promise, but which have revealed computational difficulties which tax the abilities of even the most advanced electronic computers.

The present paper has a twofold objective: first, to show a representative approach to the problem of optimum control system design and the computational problem to which such an approach leads; and second, to display a special-purpose computer which could be used to realize the optimum controller for a limited class of control problems.

## Formulation of the Problem

Consideration here is limited to discrete-time processes whose behaviour can be represented by a vector difference equation

$$x(n+1) = p[n, x(n), u(n)] \quad n = v, v+1, \dots, N-1 \quad (1)$$

where  $\mathbf{x}(n)$  is a  $k$  vector called the state of the process at the  $n$ th step and  $\mathbf{u}(n)$  is an  $r$  vector called the input. The control is to be initiated at  $n = v$  at which time the process is in state  $\mathbf{x}(v)$ . The objective of the design is to determine an algorithm for generating a control sequence  $\mathbf{u}(v), \mathbf{u}(v+1), \dots, \mathbf{u}(N-1)$  which causes the sequence of process states  $\mathbf{x}(v), \mathbf{x}(v+1), \dots, \mathbf{x}(N)$  to deviate as little as possible from a specified sequence  $\mathbf{y}(v), \mathbf{y}(v+1), \dots, \mathbf{y}(N)$ , but not at the expense of using excessively large inputs. To accomplish this objective a performance measure is defined

$$F_0 = F_0(\mathbf{e}(v), \dots, \mathbf{e}(N); \mathbf{u}(v), \dots, \mathbf{u}(N-1)) \quad (2)$$

where

$$\mathbf{e}(n) = \mathbf{x}(n) - \mathbf{y}(n) \quad (3)$$

which the optimum input sequence is to minimize. It may also be required that the process terminate at a specified state or region and this is done by imposing the condition

$$L(\mathbf{e}(N)) = 0 \quad (4)$$

Moreover, the choice of inputs is restricted since they may be physically limited in amplitude, total energy, or in some other way. These limitations are expressed in the form of  $s$  constraints

$$F_i(\mathbf{u}(v), \dots, \mathbf{u}(N-1)) \leq c_i(v) \quad i = 1, 2, \dots, s \quad (5)$$

The mathematical problem which must be solved is then to minimize the performance measure eqn (2), which is implicitly dependent only on the initial state and the input sequence, since all subsequent states depend on these, subject to the terminal condition eqn (4) and the  $s$  constraints of eqn (5). Geometrically, the set of input sequences which result in a specified value of  $F_0$  can be regarded as points on a surface in an  $(N-v)r$  dimensional space and the set of inputs which simultaneously satisfy all the constraints of eqns (4) and (5) as a region  $R$  in the space. The optimum input sequence either lies within  $R$ , in which case the constraints can be ignored, and the minimum of  $F_0$  occurs at a point where the gradient of  $F_0$  with respect to the input sequence is zero (provided the gradient exists at that point); or else it occurs on the boundary of  $R$ , in which case the boundary is tangent to the surface of minimum  $F_0$ . This tangency condition requires that the gradient of  $F_0$  be parallel and in opposite direction to the gradient of the boundary surface.

Before proceeding with the mathematical expression of these conditions, the nature of the solution will be considered. Evidently the answer to the geometric problem is a sequence  $\mathbf{u}(v), \dots, \mathbf{u}(N-1)$  which would be optimum with respect to the starting state  $\mathbf{x}(v)$ , and a corresponding optimum trajectory along which the process would move from  $\mathbf{x}(v)$  to  $\mathbf{x}(N)$  when this sequence is applied. Should a situation arise, however, where as a result of a disturbance or temporary failure, the trajectory deviates from the optimum, the remaining input sequence will no longer be optimum. The best that can be done in this case is to regard as the starting state for a new problem the state that the process is then in and to recompute a new optimum input sequence and corresponding trajectory. To allow for the possibility that such errors may occur at any step in the control interval, it is agreed that every state reached during the control interval is the starting state for a new problem, and a new optimum sequence is recomputed at each step. If this is done, however, it is not necessary to compute the entire input sequence, but only the first member thereof, since this input will

take the process to the optimum next state, at which time the computation will be repeated. It is thus clear that the algorithm by which the first input in the optimum sequence is calculated is the realization of the optimum feedback control law, while the algorithm for calculating the entire input sequence is a method of determining the optimum open-loop (preprogrammed) control.

### Derivation of Control Equations

The minimization problem formulated above is treated by the method of Lagrange multipliers. As is well known this method is capable of yielding only stationary points, not absolute minima. To avoid further difficulties it is thus assumed that the required minimum is actually attained at one of the stationary points, and is detected by a separate test if the minimum is not unique.

For the use of this technique it is convenient to write the process equations in the following form

$$\mathbf{e}(n+1) + \mathbf{y}(n+1) - \mathbf{p}(\mathbf{e}(n) - \mathbf{y}(n), \mathbf{u}(n), n) = 0 \quad (6)$$

$$n = v, \dots, N-1$$

with the initial condition  $\mathbf{e}(v) - \mathbf{x}(v) \times \mathbf{k}(v) = 0$  which is independent of the control applied. It is also convenient to write the constraints in the following form:

$$F_i - h_i = 0 \quad i = 1, 2, \dots, s \quad (7)$$

where

$$h_i \leq c_i(v) \quad (8)$$

Then the following functional is formed

$$H = H[\mathbf{e}(v), \dots, \mathbf{e}(N); \mathbf{u}(v), \dots, \mathbf{u}(N-1)]$$

$$= F_0 + \sum_{j=1}^s \lambda_j F_j + \sum_{n=v}^{N-1} (\mathbf{e}(n+1) + \mathbf{y}(n+1) - \mathbf{p}(\mathbf{e}(n) - \mathbf{y}(n), \mathbf{u}(n), n)' \mathbf{z}(n)) \quad (9)$$

where  $\mathbf{z}(v), \dots, \mathbf{z}(N-1)$  is a sequence of undetermined Lagrange multiplier vectors,  $\lambda_i$  are additional Lagrange multipliers, and the prime denotes transposition. A condition for the existence of a stationary point of  $F_0$  subject to the constraints of eqn (5) is that the sequence of vectors  $\mathbf{z}(v), \dots, \mathbf{z}(N-1)$  and the multipliers  $\lambda_i$  ( $i = 1, 2, \dots, s$ ) must be so chosen that the partial derivatives of  $H$  with respect to  $\mathbf{e}_j(n)$  and  $\mathbf{u}_i(n)$  must vanish for  $j = 1, 2, \dots, k; i = 1, 2, \dots, r; n = v, v+1, \dots, N$ . Performing the required partial differentiations with respect to the  $\mathbf{u}_i(n)$  the following set of vector equations is obtained

$$\sum_{i=0}^s \lambda_i \frac{\partial F_i}{\partial \mathbf{u}(n)} = \mathbf{P}'_u(n) \mathbf{z}(n) \quad n = v, \dots, N-1 \quad (10)$$

where  $\partial F_i / \partial \mathbf{u}(n)$  is the gradient of  $F_i$  with respect to  $\mathbf{u}(n)$ ,  $\mathbf{P}_u(n)$  is the Jacobian matrix with respect to  $\mathbf{u}(n)$  of the vector  $\mathbf{p}$  occurring in the difference eqns (1), the prime denotes transposition, and  $\lambda_0 = 1$ . (Note that the gradients of the performance measure and the constraints appear in eqn (10) in a similar manner, the difference being that the  $\lambda$  corresponding to the performance measure is fixed at unity, while the  $\lambda$ 's corresponding to the constraints are undetermined.)

Performing the partial differentiations of eqn (9) with respect to the  $e_j(n)$  yields the following set of difference equations for the vectors  $z(n)$

$$z(n-1) = P'_x(n) z(n) - \frac{\partial F_0}{\partial e(n)} \quad n = v, \dots, N-2 \quad (11)$$

where  $P_x(n)$  is the Jacobian matrix of the vector  $p$  with respect to  $x(n)$ , and  $\partial F_0 / \partial e(n)$  is the gradient of  $F_0$  with respect to  $e(n)$ . If  $e(N)$  may be varied to obtain the optimum input sequence, then the last of the sequence of vectors  $z(n)$  is given by

$$z(N-1) = - \frac{\partial F_0}{\partial e(N)} \quad (12)$$

otherwise  $z(N-1)$  is undetermined, and eqn (12) is replaced by eqn (4). The system of eqns (10) and (11) are analogous to the Euler-Lagrange equations for the optimum control of a continuous-time process which would be obtained through the use of the calculus of variations<sup>1</sup>. Equations similar to these for discrete-time processes were obtained by Kipiniak<sup>2</sup>.

In principle, eqn (10) can be solved for the sequence of inputs  $u(v), \dots, u(N-1)$  in terms of the sequence of adjoint vectors  $z(v), \dots, z(N-1)$ . Substituting the relations thus obtained into the original process eqns (1) and the adjoint eqns (11) which in general depend on the inputs through  $P_x$  would result in a system of  $2k$  difference equations with  $s$  undetermined parameters  $\lambda_1, \dots, \lambda_s$ . These, together with the initial conditions  $x(v)$  and the final condition eqn (4) or (12) and the constraints eqn (5) can be solved for the optimum adjoint vectors  $z(n)$  and sequence of states  $x(n)$ , as well as for the parameters  $\lambda_i$ . From these the optimum input sequence can be obtained, again by the use of eqn (10). (Note that if any constraint is satisfied with a strict inequality, the corresponding  $\lambda$  is equal to zero.)

If the number of steps  $M = N - v$  is appreciable the task of performing the computations indicated above is formidable and could easily surpass the capabilities of any available electronic computer. Recently several relatively effective successive approximations techniques have been studied<sup>3-5</sup> but they appear to be useful only for precomputing an optimum trajectory for a single initial state and are much too slow to be used as a feedback algorithm in real time. The next section deals with the realization of a special-purpose computer which, for a restricted class of problems, can effectively be used to perform this computation.

### A Special-purpose Computer for Terminal Control of Linear Processes

Here the construction is considered of a computer which can be used to control a linear, time-invariant process to achieve optimal terminal performance. The following three related problems are of particular concern

(1) To force the process from the initial state to a terminal state which is as close as possible to a specified state in a specified number of steps.

(2) To force the process from the initial state to a specified terminal state with minimum 'effort'.

(3) To force the process from the initial state to a specified terminal state in a minimum number of steps.

In each case the inputs which are to be used must satisfy a set of constraints in the form of eqn (5) which it will be further assumed can be expressed in the form

$$F_i^{p_i} = \sum_{n=v}^{N-1} f_i(u(n)) \quad i=1, \dots, s \quad (13)$$

where  $p_i$  is an integer  $\geq 0$ . If the  $i$ th constraint is satisfied with the equals sign then  $F_i = c_i(v)$ , and

$$\frac{\partial F_i}{\partial u(n)} = \frac{f_i^{1-p_i}}{p_i} \frac{\partial f_i}{\partial u(n)} \quad (14)$$

On the other hand, if the  $i$ th constraint is satisfied by a strict inequality, then the corresponding  $\lambda_i = 0$ . Hence eqn (10) can be written

$$\frac{\partial F_0}{\partial u(n)} + \sum_{i=1}^s \frac{\lambda_i h_i^{1-p_i}}{p_i} \frac{\partial f_i}{\partial u(n)} = P'_u(n) z(n) \quad (15)$$

For the second problem, it is assumed that the effort to be expended is one of the quantities constrained in the other two problems. In this case it is desired to minimize one of the  $F_i$ , say  $F_1$ , and eqn (15) becomes

$$\sum_{i=1}^s \frac{\lambda_i h_i^{1-p_i}}{p_i} \frac{\partial f_i}{\partial u(n)} = P'_u(n) z(n) \quad (16)$$

where  $\lambda_1 = 1$  and  $h_1$  is the minimum effort expended—which is as yet undetermined. For the first and third problems the performance measure is independent of the input sequence and thus  $\partial F_0 / \partial u(n) = 0$ . Hence eqn (16) can be used to express the relation between  $z(n)$  and  $u(n)$  for all three problems. Since  $\partial f_i / \partial u(n)$  are known (generally non-linear) functions of  $u(n)$ , eqn (16) expresses an instantaneous non-linear transformation from  $u(n)$  to  $P'_u(n) z(n)$ . This transformation is seen to depend on and is linear in the  $\lambda_i$ . Assuming that this transformation is invertible

$$u(n) = \sigma(P'_u(n) z(n); \lambda_1, \dots, \lambda_s) \quad (17)$$

where  $\sigma$  is the inverse of the transformation expressed by eqn (16). Thus it is seen that under the assumptions made the optimum input vector at the  $n$ th instant is a known transformation of the vector  $P'_u(n) z(n)$  dependent on  $s$  parameters,  $\lambda_1, \lambda_2, \dots, \lambda_s$ , which are to be determined from the constraints. In particular, the optimum initial input, which is the only one which needs to be calculated in the feedback implementation is

$$u(v) = \sigma(P'_u(v) z(v); \lambda_1, \lambda_2, \dots, \lambda_s) \quad (18)$$

The computation which must be performed is the calculation of the single vector  $z(v)$  and the parameters  $\lambda_1, \lambda_2, \dots, \lambda_s$ .

In each problem the path followed from the initial state to the terminal state is of no concern; consequently the performance measure [eqn (2)] is independent of errors along the path and the term  $\partial F_0 / \partial e(n)$  in eqn (11) is zero for  $n = v, \dots, N-2$ . Moreover, as a consequence of the linearity and time-invariance assumed of the process  $P_x = A$ , a constant matrix. Thus eqn (11) becomes

$$z(n-1) = A' z(n) \quad (19)$$

and the entire sequence of adjoint vectors can be immediately expressed in terms of the initial vector. (In a physical process  $A$  is necessarily non-singular, so the negative powers of  $A$  exist.) Specifically

$$z(n) = A'^{v-n} z(v) \quad n = v, \dots, N-1 \quad (20)$$

Moreover,  $P_u(n) = B$ , another constant matrix in a linear, time-invariant system. Thus eqn (17) can be written

$$u(n) = \sigma(B'A'^{N-n}z(v); \lambda_1, \dots, \lambda_s) \quad (21)$$

For any input sequence the state of the process at the terminal instant is given by

$$x(N) = A^{N-v}x(v) + \sum_{n=v}^{N-1} A^{N-1-n}Bu(n) \quad (22)$$

Substituting eqn (21) into (22) and rearranging terms gives

$$e(n) = A^{N-v}[\bar{e}(N, v) + A^{-1}g_{N-v}(z(v); \lambda_1, \dots, \lambda_s)] \quad (23)$$

where

$$\bar{e}(N, v) = e(v) + y(v) - A^{v-N}y(n) \quad (24)$$

can be called the 'effective error vector' (so called because it equals the observed error when the desired trajectory is a solution to the unforced equations; if the observed error  $e(v)$  is used in place of  $\bar{e}(v)$ , this is equivalent to extrapolating the desired trajectory by a solution of the unforced equations), and the vector

$$g_n(z(v); \lambda_1, \dots, \lambda_s) = \sum_{m=0}^{n-1} A^{-m}B\sigma(B'A'^{-m}z(v); \lambda_1, \dots, \lambda_s) \quad (25)$$

is a calculable non-linear vector-valued function of  $z(v)$ ,  $N-v$ , and  $\lambda_i$  ( $i = 1, 2, \dots, s$ ).

The relationship between the error  $e(v)$ ,  $z(v)$ , and  $e(N)$  is thus completely expressed by eqn (23). To obtain the relationship between  $e(v)$  and  $z(v)$ , which determines the input to be applied at  $n = v$ , it is necessary to eliminate  $e(N)$  from eqn (23) and to evaluate the  $\lambda_i$  from the constraint relations of eqn (5).

For the first problem, it is assumed that the terminal error is measured by a positive-definite function  $L(e(N))$  which is thus the performance measure  $F_0$ . Hence, from eqn (12)

$$z(N-1) = A'^{v-N+1}z(v) = -\phi(e(N)) \quad (26)$$

where  $\phi(e(N)) = \partial L / \partial e(N)$ . Thus, from eqns (26) and (23)

$$-A'^{N-v-1}\phi[A^{N-v-1}(A\bar{e} + g_{N-v})] = z(v) \quad (27)$$

where it is understood that  $g_{N-v} = g_{N-v}(z(v); \lambda_1, \dots, \lambda_s)$ . The solution of eqn (27) and the determination of the  $\lambda_i$  from the constraints gives the optimum  $z(v)$ , which, in accordance with eqn (18), is transformed to the optimum input  $u(v)$ . This process is repeated at the next step with  $N$  replaced by  $N-1$ .

A special-purpose analogue computer for performing the computation required by eqn (27) is shown in Figure 1, in which each box represents a linear or non-linear instantaneous transformation of a vector of signals into another vector. In particular  $A$  is a constant linear transformation,  $A^{N-v-1}$  and  $A'^{N-v-1}$  are time-varying linear transformations,  $\phi$  and  $\sigma$  are constant non-linear transformations determined by eqns (26) and (18), respectively, and  $g$  is a time-varying non-linear transformation determined by eqn (25). The  $g$  unit has parameters  $\lambda_i$  which must be set in accordance with the constraints. Since these parameters completely determine the  $g$  unit and the  $\tau$  unit, and the other units are determined, the only computation remaining is the determination of the correct settings of  $\lambda_i$ .

The computer is operated by setting the number of steps to go  $M = N-v$  into the time-varying units and adjusting the  $\lambda_i$  such that the constraints are satisfied and  $L(e(N))$  is minimized.

The terminal error which would result for a given setting of the  $\lambda_i$  appears explicitly to the left of the  $\phi$  unit and can thus be brought into the subsidiary computer for the determination of the  $\lambda_i$ . It is also necessary to know whether a given setting of the  $\lambda_i$  will satisfy the constraints on the  $u(n)$ . A realization of the  $g$  unit which places the entire input sequence in evidence is shown in Figure 2. The unit consists of a variable number of sections and a switching arrangement (not shown) for connecting

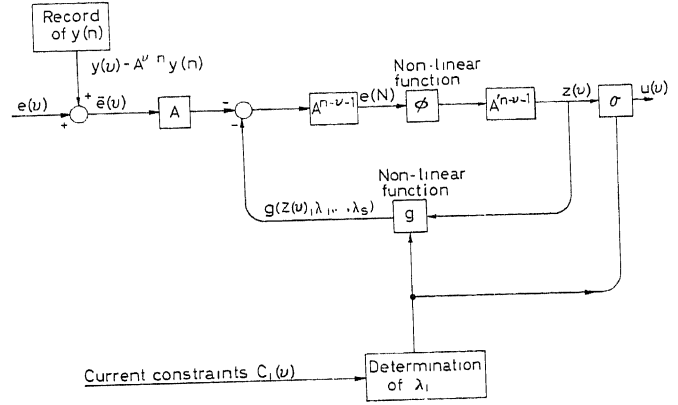


Figure 1. Structure of optimum controller

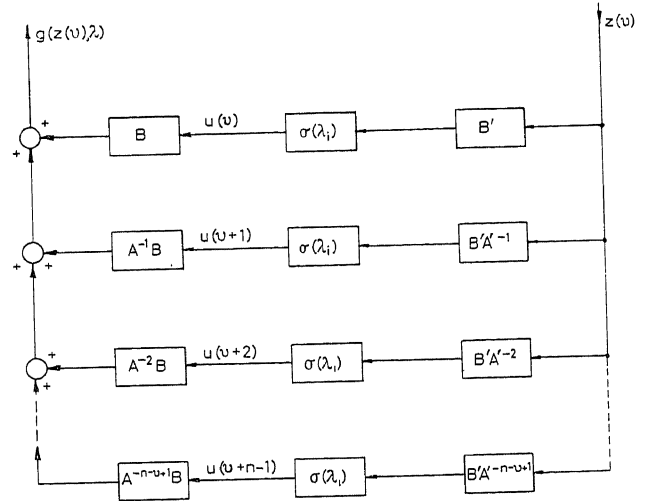


Figure 2. Structure of  $\sigma$  unit

as many sections as there are steps to go. It is seen that the inputs which would be generated by a given setting of the  $\lambda_i$  appear explicitly at the outputs of  $\sigma$  units. (Note also that the optimum present input can be taken from the first  $\sigma$  unit, eliminating the need for the  $\sigma$  unit of Figure 1.)

A useful class of constraints is represented by the general form

$$F_i^{p_i} = \sum_{n=v}^{N-1} \sum_{j=1}^r |u_j(n)|^{p_i} = h_i^{p_i-1} \quad (28)$$

which is essentially a norm of the sequence  $u(v), \dots, u(N-1)$  in an  $(N-v)r$  dimensional space. When  $p_i = 2$  the constraint is on the total available energy. When  $p_i = 1$ , eqn (28) becomes

$$F_i = \sum_{n=v}^{N-1} \sum_{j=1}^r |u_j(n)|$$

which is a constraint on the total effort, and is useful in connection with control of space vehicles with limited thrust capability. As  $p_i \rightarrow \infty$ ,  $F_i \rightarrow \sup_{n,j} |u_j(n)|$ , the familiar amplitude constraint.

The partial derivative of  $F_i$  with respect to  $u_j(n)$  is given by

$$\frac{\partial F_i}{\partial u_j(n)} = \left| \frac{u_j(n)}{h_i} \right|^{p_i-1} \text{sgn } u_j(n) \quad (29)$$

where

$$\text{sgn } u = \begin{cases} +1 & u > 0 \\ 0 & u = 0 \\ -1 & u < 0 \end{cases}$$

Thus each component of  $\partial F_i / \partial \mathbf{u}(n)$  is a non-linear transformation of the corresponding component of  $\mathbf{u}(n)$  as given by eqn (29).

If all the constraints are of this form, then the  $j$ th component of the vector  $B'z(n)$  in eqn (16) can be written

$$[B'z(n)]_j = \sum_{i=1}^s \lambda_i \left| \frac{u_j(n)}{h_i} \right|^{p_i-1} \text{sgn } u_j(n) \quad j=1, 2, \dots, k \quad (30)$$

Thus, for  $\lambda_i \geq 0$ , and at least one  $\lambda_i \neq 0$ ,  $[B'z(n)]_j$  is a non-decreasing symmetric function of  $u_j(n)$ . Thus it is evident that  $u_j(n)$  is likewise a non-decreasing, symmetric function of  $[B'z(n)]_j$ . As an illustration, select  $p_1 = 1$ ,  $p_2 = 2$ ,  $p_3 = \infty$ . Then the graph of  $[B'z(n)]_j$  vs.  $u_j$  has the appearance of Figure 3(b) and  $u_j$  vs.  $[B'z(n)]_j$  has the appearance of Figure 3(c). The non-linear characteristic  $\sigma$  is seen to have three regions: a dead zone resulting from the total effort constraint, a linear region resulting from the quadratic constraint, and a saturation region resulting from the amplitude constraint. (As shown in Figure 3(c) the non-linear characteristic is single valued and no difficulty would result in using this characteristic in the analogue computer. If the quadratic constraint were absent, however, there would be discontinuities at  $\pm \lambda_1$  and we could expect difficulties to arise. To avoid such difficulties it may be permissible to assume a small but finite linear region, even when there is no quadratic constraint, since a quadratic constraint may always

be assumed to exist together with either a total effort or an amplitude constraint.) In this case  $\lambda_3$  is automatically taken into account, and the adjustments required are in the width of the dead zone and the slope of the linear region. Note that as the slope of the linear region becomes infinite (i.e., the energy constraint becomes negligible) then every input  $u_j(n)$  is either zero or the maximum value  $h_3$ .

If there is only a single constraint in the form of eqn (28) to be accounted for, then the adjustment of the corresponding single parameter  $\lambda$  is considerably simplified. It is observed that if the dot product of the vector  $\mathbf{z}$  with the vector  $\mathbf{g}$  is taken, one obtains

$$\mathbf{z}'(v) \mathbf{g}_{N-v}(\mathbf{z}(v)) = \sum_{n=0}^{N-1-v} \mathbf{z}'(v) A^{-n} B \sigma(B' A'^{-n} \mathbf{z}(v)) \quad (31)$$

but  $\sigma(B' A'^{-n} \mathbf{z}(v)) = \mathbf{u}(v-n)$ . Consequently, eqn (31) can be written

$$\begin{aligned} \mathbf{z}'(v) \mathbf{g}_{N-v}(\mathbf{z}(v)) &= \sum_{n=0}^{N-1-v} \sigma^{-1}(\mathbf{u}(v-n))' \mathbf{u}(v-n) \\ &= \sum_{n=v}^{N-1} \mathbf{u}'(n) \sigma^{-1}(\mathbf{u}(n)) = \sum_{n=v}^{N-1} \sum_{j=1}^r u_j(n) \sigma^{-1}(u_j(n)) \end{aligned}$$

But  $\sigma^{-1}(u_j(n)) = \lambda |u_j(n)/h^{p-1}| \text{sgn } u_j$ . Whence

$$\mathbf{z}'(v) \mathbf{g}_{N-v}(\mathbf{z}(v)) = \lambda \sum_{n=v}^{N-1} \sum_{j=1}^r |u_j| p / h^{p-1} = \lambda h$$

Thus the value of  $h$  which is obtained for a trial value of  $\lambda$  is given by the dot product of  $\mathbf{z}$  with  $\mathbf{g}$  divided by  $\lambda$ :

$$h = \frac{\mathbf{z}'(v) \mathbf{g}_{N-v}(\mathbf{z}(v))}{\lambda} \quad (32)$$

It is thus necessary only to evaluate this dot product to determine whether the constraint is satisfied. To avoid the necessity of adjusting the  $\sigma$  unit in accordance with the undetermined  $h$ ,  $\lambda$  can be replaced by a parameter  $k = \lambda/h^{p-1}$  ( $p < \infty$ ). In this case the value of  $h$  obtained for a given setting of  $k$  is

$$h = \left[ \frac{\mathbf{z}'(v) \mathbf{g}_{N-v}(\mathbf{z}(v))}{k} \right]^{1/p} \quad (33)$$

For a single constraint the adjustment of  $k$  can be accomplished by setting  $k$  to a very large value (which is seen to make the transmission through the  $g$  unit and likewise  $g'z$  small) and sweeping  $k$  towards a very small value (thereby increasing the transmission through  $g$ ) until the right-hand side of eqn (33) is exactly equal to the available quantity  $c(v)$ . Thus the subsidiary computer needs only to evaluate  $h$  from eqn (33) and compare this value with  $c(v)$  to obtain the correct setting of  $k$ . This technique could be used as an approximate realization for an amplitude constraint by making  $p$  a large but finite value<sup>6</sup>.

For multiple constraints the same adjustment procedure can be followed, except that each of the  $k_i = \lambda_i/h_i^{p_i-1}$  have to be adjusted in sequences of small steps and the calculation of  $h_i$  cannot be simplified to the evaluation of a dot product, unless each constraint applies to components of the input vector which do not appear in the other constraints, in which case the dot product technique remains applicable. See Friedland<sup>7</sup>, for example.

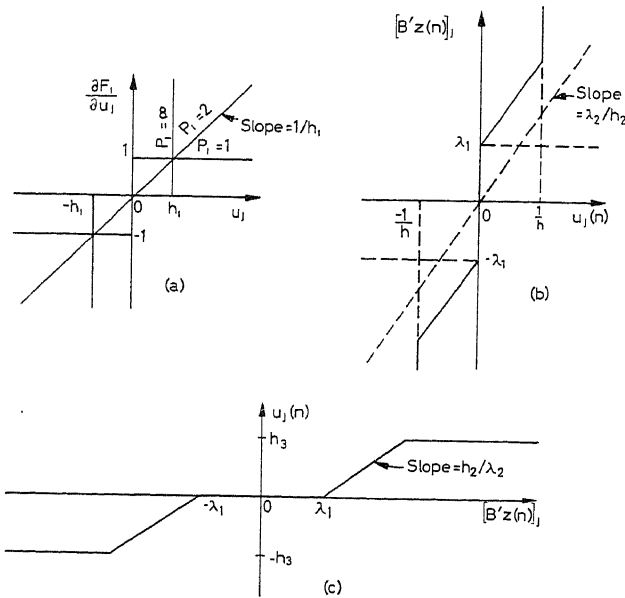


Figure 3. Determination of non-linear characteristic for  $\sigma$  unit

Consider now the second problem, that of minimizing the total effort required to reduce the terminal error to zero (assuming this can be done) in a specified number of steps to go. In this case it is required that  $e(N)$  in eqn (23) be zero, and this establishes the following relationship between  $\bar{e}$  and  $z$

$$-A\bar{e}(N, v) = g_{N-v}(z(v), \lambda_1, \dots, \lambda_s) \quad (34)$$

To solve this equation for  $z(v)$  entails the inversion on the non-linear function  $g$ . This is accomplished with any accuracy required in the analogue computer by making the non-linear function  $\phi$  have a very large (ideally infinite) transmission. This is equivalent to using a terminal error criterion  $L(e(N))$  which imposes very large penalties for terminal errors and resembles the 'penalty function' approach suggested by Kelley<sup>3</sup>. The adjustment procedure is similar to that described for the problem of minimizing the terminal error. In fact, the computer has a built in flexibility which permits its use for either problem. The adjustment can be started with the assumption that the error cannot be reduced to zero. Set the transmission of  $\phi$  to a finite level and then adjust the parameters of the  $g$  unit in an attempt to satisfy the constraints. If this assumption is warranted, then at least one of the  $\lambda_i$  (or  $k_i = \lambda_i/h_i^{p_i-1}$ ) will be non-zero, since the optimum will result in at least one constraint being satisfied with the equals sign. However, if this assumption is not warranted, the implication is that all of the constraints are satisfied by the strict inequality and all the  $\lambda_i$  will be zero. In this case the transmission through the  $g$  unit will be infinite, the error will automatically be reduced to zero, and the effort  $F_1$  will simultaneously be minimized, since the computer will then be solving eqn (34). In practice, of course, the transmission of the  $g$  unit cannot be permitted to be infinite, but it can be made large enough to ensure that the terminal error is below any tolerable level.

For the third problem, that of minimizing the number of steps to reduce the error to zero, use as the performance measure,  $F_0$ , the number of steps that the error is non-zero. Since  $\partial F_0 / \partial e(n) = 0$ ,  $n = v, \dots, N$ , the optimum control law is also given by eqn (34); but here  $N$  is the smallest integer for which all the constraints are satisfied. It is noted that at the optimum  $N$  all the constraints could be satisfied by strict inequalities, although for any smaller  $N$  it is not possible to reduce the error to zero with any input which satisfies the constraints.

The analogue computer could be used to minimize the number of steps by starting with a number of steps to go  $M \geq M_{\min}$  where  $M_{\min}$  is the smallest number of steps in which the error could be reduced to zero in the absence of any constraints and is equal to the smallest integer  $\mu$  for which the rank of  $[B \ AB \dots A^{\mu-1}B]$  is  $k$ . (It has been shown that  $\mu \leq k$ , the order of the system unless the process is not controllable, in which case no such  $\mu$  exists<sup>8</sup>.) At this setting of  $M$  it can be determined whether all the constraints are satisfied with strict inequalities, in which case the transmission through  $g$  is infinite and the error is reduced to zero. If one of the constraints cannot be satisfied,  $M$  is increased by one and the process repeated until the value of  $M$  reaches results in the satisfaction of all the constraints.

### Recapitulation and Conclusions

A straightforward calculus development has been used to obtain a system of algebraic equations which must be solved

in order to determine optimum input sequence for a discrete-time process. If this process is to be controlled by the feedback method, wherein the present input is to be calculated in terms of the present error and available 'effort', then this computation must be performed in real time. A special-purpose computer has been displayed which it is feasible to use for implementing the required computation. The essential feature of this computer is the use of a feedback arrangement to solve a system of non-linear equations. Each component of the computer is ideally instantaneous, so that in principle the computation for a given setting of the parameter  $\lambda_i$  is instantaneous. In practice, one will have to wait until the transient, which will inevitably result owing to the presence of parasitic dynamic elements, decays sufficiently before its output can be observed. Moreover, the computer will require careful design to prevent the occurrence of unstable transients which could result when the transmission through the various units is high.

Nevertheless, the analogue computer appears naturally suited to the solution of the problem, since there is no systematic method of obtaining an exact solution to eqn (27) or to eqn (34) on a digital computer. A digital computer could be used to obtain an iterative solution to eqn (27). A suitable iteration rule would be

$$z^{(t+1)} = z^{(t)} + \psi \{z^{(t)} + A^{N-v-1} \phi [A^{N-v-1}(\bar{A}e + g(z^{(t)}))]\} \quad (35)$$

where  $t$  indicates the iteration number. The function  $\psi$  must be selected to ensure convergence of  $z^{(t)}$ , and, from a practical standpoint, that the convergence be rapid. The computation performed by such a computation technique turns out to be quite similar to the steepest descent techniques referred to earlier, and could probably not be performed in real time, unless the starting vector  $z^{(0)}$  is close to the optimum.

It is noted that the determination of the settings of the  $\lambda_i$  do not change along an optimum trajectory, and in principle, could be left fixed after their initial determination. However, the use of feedback implies that errors will cause the process to deviate from the initially computed optimum trajectory, and the  $\lambda_i$  can accordingly be expected to change from step to step. But, if the deviation from optimum path is small, the adjustments required to obtain the correct setting of the  $\lambda_i$  can also be expected to be small, and whatever technique is used to obtain the initial adjustments will work much more rapidly for subsequent adjustments. As an approximation the  $\lambda_i$  could be fixed at their initial value and thereby greatly simplify the resulting computer. Such a simplification might be warranted in aerospace vehicles where the computer which makes the initial determination of the  $\lambda_i$  could be detached from the vehicle at launching.

### References

- 1 FRIEDLAND, B. The structure of optimum control systems. *J. Basic Engng, Trans. Amer. Soc. mech. Engrs* 84 (1962) 1
- 2 KIPINIAK, W. *Dynamic Optimization and Control: A Variational Approach*. 1961. Massachusetts Institute of Technology Press and Wiley
- 3 KELLEY, H. Method of gradients. *Optimization Techniques*. Chap. 6 (Ed.) G. Leitman, 1962. New York; Academic Press
- 4 BRYSON, A. E. and DENHAM, W. F. A steepest descent method for solving optimum programming problems. *J. appl. Mech. Trans. Amer. Soc. mech. Engrs* 29 (1962) 247

- <sup>5</sup> HO, Y. C. A successive approximation technique for optimal control systems subject to input saturation. *J. Basic Engng, Trans. Amer. Soc. mech. Engrs* 84 (1962) 33
- <sup>6</sup> KIRILLOVA, F. M. A limiting process in the solution of an optimum control problem. *Appl. Math. Mech., Leningr.* 24 (1960) 398
- <sup>7</sup> FRIEDLAND, B. The design of optimum controllers for linear processes with energy limitations. *J. Basic Engng, Trans. Amer. Soc. mech. Engrs* (1963) Pap. No. 62-JACC-8
- <sup>8</sup> KALMAN, R. E. On the general theory of control systems. *Automatic and Remote Control*. Vol. 1 p. 481 (1961) London; Butterworths

# DISCUSSION

J. L. SHEARER, *Penn State University, University Park, Pa., U.S.A.*

Have you worked out a block diagram structure for the  $\sigma$  unit?

B. FRIEDLAND, *in reply*

As indicated in the text, the  $\sigma$  unit is the non-linear device which transforms  $P_u'(n)Z(n)$  into  $u(n)$  and is obtained by solving eqn. (16) for  $u(n)$ . An example of this calculation is given following eqn. (30) and is accompanied by the graphical illustrations of *Figure 3*.



# Theory of Minimum Time Discrete Regulators

C. A. DESOER, E. POLAK and J. WING

## Summary

This paper unifies and extends the ideas and techniques used by the authors in relation with the minimal time regulator problem. The system considered is a multiple input system described by a differential equation of the form  $\dot{x} = F(x) + Bu(t)$  where  $u$  is either a PAM or PWM bounded control vector. This paper consists of five sections. In the first, the problem is carefully stated in a suitable general setting, and the concept of optimal strategy is defined; in the second, an optimal strategy for a multiple input PAM system with a completely controllable linear plant is described fully, and a special case is discussed; in the third, the concept of equivalent optimal strategies is defined, and sufficient conditions for equivalence are indicated. In the fourth section equivalence is used to derive the optimal strategy of a PWM system whose linear plant has real and distinct eigenvalues, and finally in the last section, some extensions are mentioned.

## Sommaire

Cette communication présente une synthèse et une extension des idées et des techniques préconisées par les auteurs pour la solution du problème du régulateur à temps minimal. Le système considéré est un système à entrées multiples régi par une équation différentielle de la forme  $\dot{x} = F(x) + Bu(t)$  où  $u(t)$  est un vecteur de commande borné, à modulation soit par amplitude d'impulsions à entrées multiples avec une installation linéaire complètement réglable et on discute d'un cas spécial. Dans la troisième, on définit la notion de stratégies optimales équivalentes et on énumère les conditions suffisantes d'équivalence. Dans la quatrième partie, on utilise la notion d'équivalence pour établir la stratégie optimale d'un système à modulation par largeur d'impulsions dont l'installation linéaire possède des valeurs propres réelles et distinctes. Enfin, dans la dernière partie, on indique quelques extensions.

## Zusammenfassung

Der Aufsatz gibt eine Vereinheitlichung sowie Erweiterung der Gedanken und Verfahren, die die Autoren im Zusammenhang mit der schnelligkeitsoptimalen Regelung verwendeten. Das betrachtete System hat mehrere Eingänge; es wird durch eine Differentialgleichung der Form  $\dot{x} = F(x) + Bu(t)$  beschrieben, wobei der Vektor  $u$  die pulsamplituden- oder pulsbreiten-modulierte Regelung mit Sättigung darstellt. Der Aufsatz besteht aus fünf Abschnitten. Der erste Abschnitt enthält das sorgfältig dargestellte Problem in einer geeigneten allgemeinen Fassung und die Definition des Begriffes der optimalen Strategie. Der zweite Abschnitt beschreibt eingehend die optimale Strategie für ein pulsamplituden-moduliertes System mit mehreren Eingängen bei völlig regelbarer linearer Regelstrecke; ein Sonderfall wird betrachtet. Im dritten Abschnitt wird der Begriff der äquivalenten optimalen Strategie festgelegt und ein Hinweis auf hinreichende Bedingungen für die Äquivalenz gegeben. Der Begriff der Äquivalenz wird im vierten Abschnitt dazu benutzt, die optimale Strategie für ein System mit Pulsbreiten-Modulation abzuleiten, dessen lineare Regelstrecke reelle und verschiedene Eigenwerte hat. Schließlich sind im letzten Abschnitt einige Erweiterungsmöglichkeiten erwähnt.

## Introduction

One of the fundamental problems of engineering is that of improving the performance of a given system. In the early days of automatic control, this consisted almost exclusively of compensation techniques. Under the impact of decision theory and Wiener's<sup>1</sup> and Kolmogorov's<sup>2</sup> work on the minimization of mean squared error, strong interest developed in the mathematical formulation of optimization problems. Later McDonald<sup>3</sup> emphasized the importance of the constraints on the controls, formulated the minimum time problem, and introduced the idea of switching curve. It is now customary to state an optimum control problem in the following way. Given a system described by a set of differential or difference equations, a restricted set of (admissible) controls, and the requirement that the system be transferred from its initial state to a set in its state space, find an optimal strategy such that it specifies an admissible control fulfilling the requirement and minimizing a cost.

In the last ten years a large number of papers appeared in the field of optimal control. They have culminated in a number of important results such as Bellman's principle of optimality<sup>4</sup>, Pontryagin's maximum principle<sup>5</sup>, the bang-bang principle emphasized by La Salle<sup>6-8</sup>, the concept of controllability<sup>9</sup> and canonical structure<sup>10</sup>. At the same time other investigators sought to implement the ideas of optimal control by constructing optimal switching surfaces<sup>11, 12</sup>, digitizing the maximum principle<sup>13</sup>, devising various iterative techniques<sup>14, 15</sup>, and using the calculus of variations<sup>8, 16, 17</sup>.

In parallel with the work on continuous systems, rapid developments occurred in discrete systems. Essentially there are two approaches to such problems. In the first Krasovskii<sup>18, 19</sup> and Tsypkin<sup>20</sup> showed how the problem could be handled either by a reduction to the Krein  $L$ -problem or else by the maximization of a functional by the calculus of variations. The second approach is analogous to techniques used on mazes and discrete state machines<sup>21</sup> and to dynamic programming. This method consists of partitioning the state space of the system into sets  $R_N$ , ( $N = 1, 2, \dots$ ), from which a transition to the desired state is possible in  $N$  steps and no fewer. The minimal step paths, hence the optimal controls associated with these, connecting the initial state with the desired state can then be found either by inspection or by making use of the algorithms used in the construction of the sets  $R_N$ . Consequently, these can be used to define the optimal strategy.

This method was used by Kalman<sup>22</sup> to consider a minimal time problem of saturating PAM discrete regulator systems. Later Nelson<sup>23</sup> considered a similar problem for the PWM case. The authors of this paper considered the same problem as Kalman, introduced the concept of critical surface, expressed the optimal strategy in terms of 'distance' measurements to this critical surface, and showed the relationship between the discrete case and the continuous case<sup>24</sup>. These early results were generalized and extended to more complicated situations<sup>25-27</sup> and also applied to PWM systems<sup>28</sup>. An important new idea was the

introduction of an equivalence relation between systems<sup>24</sup>. This equivalence can be used to construct an optimal strategy for one system from the known one of another. This technique has been successfully applied to obtain the optimum control for PWM systems<sup>24</sup>, and for systems with non-linear plants<sup>31, 32</sup>.

This paper considers from a unified point of view and extends recent results pertaining to the minimal time regulator problem. The paper consists of five sections. In the first section the problem is carefully stated in a suitable general setting, and the concept of optimal strategy is defined. An optimal strategy for a multiple input PAM system with a completely controllable linear plant is described fully in the second section, and a special case is discussed. In the third, the concept of equivalent optimal strategies is defined, and sufficient conditions for equivalence are indicated. In the fourth section, equivalence is used to derive the optimal strategy of a PWM system whose linear plant has real and distinct eigenvalues, and finally in the last section, some extensions are mentioned.

### Statement of the Problem

All the systems considered in this paper have a block diagram representation shown on Figure 1. In each case the system consists of a plant, a modulator, and a computer.

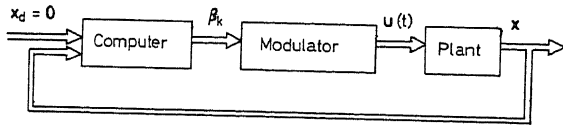


Figure 1. Block diagram of a representative system

The plant is either a linear or a non-linear time invariant dynamical system<sup>9, 10</sup> with  $m$  input channels, which is completely controllable<sup>9\*</sup> with respect to  $\mathbf{0}$  by the modulator output and is described by a differential equation whose solution is unique with respect to initial conditions and inputs. In general, the differential equation has the form

$$\dot{\mathbf{x}} = \mathbf{F}(\mathbf{x}, \mathbf{u}) \quad (1)$$

where  $\mathbf{x} = \text{col}(x_1, x_2, \dots, x_n)$  is the state vector,  $\mathbf{u}(t) = \text{col}[u_1(t), u_2(t), \dots, u_m(t)]$  is the  $m$ -channel modulator output and  $\mathbf{F}$  is a  $n \times 1$  vector valued function, Lipschitzian with respect to  $\mathbf{x}$  and such that  $\mathbf{F}(\mathbf{0}, \mathbf{0}) = \mathbf{0}^{80}$ . In the particular case when the plant is a linear dynamical system, the differential equation takes on the form:

$$\dot{\mathbf{x}} = \mathbf{B}\mathbf{x} + \mathbf{E}\mathbf{u}(t) \quad (2)$$

where  $\mathbf{B}$  is a  $n \times n$  constant matrix and  $\mathbf{E}$  is a constant  $n \times m$  control matrix with columns  $\mathbf{e}_1, \mathbf{e}_2, \dots, \mathbf{e}_m$ . The condition of complete controllability imposes additional restrictions on the plant, which must be worked out for each particular case.

The modulator may be either a pulse amplitude (PAM), or a pulse width (PWM) modulator. When it is necessary to distinguish between the modulators, a PAM output is designated by  $\mathbf{u}^a$  and a PWM output is designated by  $\mathbf{u}^w$ ; otherwise the

symbol  $\mathbf{u}$  will be used for either modulator output. The modulation laws for the two modulators are given as follows:

Let in both cases  $\zeta_k^i = \pm 1$ ,  $0 \leq \alpha_k^i \leq 1$  for all  $i$  and  $k$

$$\text{PAM: } u_i^a(t) = \zeta_k^i \alpha_k^i M \text{ for } (k-1)T \leq t < kT \quad (3)$$

$$\text{PWM: } u_i^w(t) = \begin{cases} \zeta_k^i M & \text{for } (k-1)T \leq t < (k-1+\alpha_k^i)T \\ 0 & \text{for } (k-1+\alpha_k^i)T \leq t < kT \end{cases} \quad (4)$$

where in both cases,  $k = 1, 2, 3, \dots$ ;  $i = 1, 2, \dots, m$ .  $M$  is the maximum amplitude of the modulator pulse output,  $T$  is the sampling period. Since, during any sampling period, the modulator output is characterized by  $m$  numbers, let the vector  $\beta_k \triangleq \text{col}(\zeta_k^1 \alpha_k^1, \zeta_k^2 \alpha_k^2, \dots, \zeta_k^m \alpha_k^m)$  be introduced to define the modulator output during the interval  $(k-1)T \leq t < kT$ . Throughout the paper, for all  $i$ 's and  $k$ 's,  $\zeta_k^i = \pm 1$  and  $0 \leq \alpha_k^i \leq 1$ . Any control defined by such a vector and by either modulator law (3) or (4) is called admissible. Finally, let  $\mathbf{u}_{[0, NT]}$  be a vector-valued function equal to the control  $\mathbf{u}(t)$  for  $0 \leq t \leq NT$  and equal to  $\mathbf{0}$  anywhere else. It is important to note that there is a one-to-one correspondence between  $\mathbf{u}_{[0, NT]}$  and the sequence  $\{\beta_k\}_{k=1}^N$ . The set of all finite sequences  $\{\beta_k\}_{k=1}^N$  where the components of each  $\beta_k$  satisfy the constraints specified above is called the sequence space  $S$ .

The computer generates the  $\beta_k$ 's at  $t = (k-1)T$ ,  $k = 1, 2, \dots$ , on the basis of  $\mathbf{x}_{k-1}$ , the observed value of the state at that instant, in such a manner as to transfer the system from  $\mathbf{x}_{k-1}$  to the desired state  $\mathbf{0}$  in the smallest number of sampling periods possible. Let this number be  $N - k + 1$ . It may be assumed for the purpose of the following that the computer generates the whole vector sequence  $\{\beta_j\}_{j=k-1}^N$  at  $t = (k-1)T$ , feeds the vector  $\beta_k$  to the modulator and stores the rest of the sequence. At  $t = kT$ , the computer feeds  $\beta_{k+1}$  to the modulator, etc. In practice, the computer needs only to compute the first element of the sequence at each sampling period.

For the systems under consideration, an optimal strategy is a function,  $f$ , whose domain is the state space  $E^n$  and whose range is the sequence space  $S$ , with the following two properties. First, for any  $\mathbf{x}$  in  $E^n$ ,  $f(\mathbf{x})$  is a finite sequence, say of length  $N$ , such that the modulator output defined by the sequence  $f(\mathbf{x})$  transfers the state  $\mathbf{x}$  to the zero state,  $\mathbf{0}$ , in  $N$  sampling periods; and second, there is no sequence of length smaller than  $N$  which defines a modulator output capable of transferring  $\mathbf{x}$  to  $\mathbf{0}$  in less than  $N$  sampling periods.

The minimal time regulator problem can be stated as follows. Given a completely controllable system such as the one shown on Figure 1, together with its plant dynamics and its modulator law, find an optimal strategy for that system.

For any integer  $N$ ,  $R_N$  is the set of all initial states that can be transferred to the origin by an admissible control in  $N$  sampling periods and no fewer. Let  $R_N' = \bigcup_{k=0}^N R_k$ , i.e., it is the set of

all states that can be transferred to the origin by an admissible control in no more than  $N$  sampling periods. The sets  $R_N$  can be used to characterize an optimal strategy. A mapping  $f$  of the state space  $E^n$  into the sequence space  $S$  is an optimal strategy if and only if, for all  $N$ , and, for all  $\mathbf{x} \in R_N$ , the first element of the sequence  $f(\mathbf{x})$ , say  $\beta_1$ , defines a modulator output,  $\mathbf{u}_{[0, T]} = \hat{\mathbf{u}}(\beta_1)$ , which takes the state  $\mathbf{x}$  at time  $0$  to a state  $\mathbf{y}$  at time  $T$  such that  $\mathbf{y} \in R_{N-1}$ .  $\mathbf{u}_{[0, T]} = \hat{\mathbf{u}}(\beta_1)$  is said to be optimal.

\* A system is said to be completely controllable with respect to the origin if for any initial state there exists an admissible control which steers this state to the origin in finite time.

### An Optimal Strategy for a PAM System with Linear Plant

Suppose the plant is a linear, time invariant, dynamical system and suppose its state equation is given by eqn (2). To avoid trivialities, the rank of  $E$  is assumed to be  $m$ . Let the control  $u^a(t)$  be of the form (3). In view of the form of the modulation law, the system in terms of its state transition equation is now described. Let  $x_k$  be the state at time  $kT$ , then for  $k = 0, 1, 2, \dots$

$$x_{k+1} = Ax_k + \sum_{j=1}^m \zeta_{k+1}^j \alpha_{k+1}^j d_j(T) \quad (5)$$

$$\text{where } A = \exp BT, \text{ and } d_j(t) = \int_0^t M e^{B(t-s)} e_j ds$$

$$(j=1, 2, \dots, m).$$

Iterating (5)  $N$  times, one obtains

$$x_N = A^N x_0 + \sum_{k=1}^N \sum_{j=1}^m \zeta_k^j \alpha_k^j A^{N-k} d_j(T) \quad (6)$$

If one defines

$$r_k(t) = -A^{-k} d_j(t) \quad (j=1, 2, \dots, m; k=1, 2, \dots) \quad (7)$$

then, if  $x_N = 0$ , one obtains\*

$$x_0 = \sum_{k=1}^N \sum_{j=1}^m \zeta_k^j \alpha_k^j r_k^j \quad (8)$$

An optimal strategy for such a PAM system is now described. First, controllability is considered<sup>26</sup>.

**Theorem 1**—The system (2) with modulation law (3) is completely controllable by admissible controls if and only if the set of  $mn$  vectors  $\Sigma_{mn} = \{r_k^j, 1 \leq k \leq n, 1 \leq j \leq m\}$  spans the state space  $E^n$  and the eigenvalues of  $A$  lie in the closed unit circle  $|\lambda| \leq 1$ .

If one eigenvalue of  $A$ , say  $\lambda_1$ , lies outside the unit circle, then a state  $x$  is completely controllable by admissible controls provided the projection of  $x$  on the invariant subspace associated with  $\lambda_1$  lies within an open set which is determined by  $A$  and  $E$ .

The optimal strategy can be described in terms of  $m$  hypersurfaces  $\mathcal{C}^j$  called the critical hypersurfaces<sup>26, 27</sup>. With the  $j$ th component of the control vector, ( $1 \leq j \leq m$ ), is associated a critical hypersurface  $\mathcal{C}^j$ . The critical hypersurfaces satisfy the following intersection property. For  $j = 1, 2, \dots, m$ , any straight line parallel to  $r_1^j$  intersects  $\mathcal{C}^j$  at one and only one point. Therefore given any state  $x$ , the  $m$  numbers  $\lambda_j(x)$  defined by the relation  $x = \lambda_j(x) r_1^j + y_j$  where  $y_j \in \mathcal{C}^j$  are uniquely defined. Geometrically  $\lambda_j(x)$  measures the 'distance' between  $x$  and  $\mathcal{C}^j$ . Provided 'distance' is measured in the direction of  $r_1^j$  and the Euclidian length  $\|r_1^j\|$  is taken as the unit of length. If  $\text{sat } \gamma$  is defined as  $\gamma/\|\gamma\|$  whenever the real number  $\gamma$  is outside  $[-1, 1]$  and  $\text{sat } \gamma = \gamma$  is put for all other  $\gamma$ , then the vector  $\phi(x) = \text{col}[\text{sat } \lambda_1(x), \dots, \text{sat } \lambda_m(x)]$  defines an admissible control, for that  $x$ . Let  $D$  be the  $n \times m$  matrix whose columns are the vectors  $d_j(T)$ ,  $j = 1, 2, \dots, m$ . It can be shown<sup>27</sup> that if  $x_0 \in R_N$  at  $t = 0$ ,

then  $Ax_0 + D\phi(x) \in R_{N-1}$  at  $t = T$ . In other words, the control defined over  $[0, T]$  by the vector  $\phi(x)$ , i.e.  $u[\phi(x)]$ , is optimal. Since this holds for all  $n$  and all  $x_0$  in  $R_N$ , the function  $\phi$  can be used to define an optimal sequence for  $x_0$ , namely  $\{\phi(x_0), \phi(x_1), \dots, \phi(x_{N-1})\}$  where  $x_{k+1} = Ax_k + D\phi(x_k)$ . In other words, the sequence  $\{\phi(x_0), \phi(x_1), \dots, \phi(x_{N-1})\}$  is the value taken at  $x_0$ , namely  $f(x_0)$ , by an optimal strategy  $f$ . Going back to the block diagram of Figure 1, it is seen that the function  $\phi$  specifies an optimal programme for the computer. The process of evaluating  $\phi(x)$  by 'distance' measurements at the beginning of each sampling period is illustrated by Figure 2. It shows the state space of a second order system with two inputs; the critical surfaces  $\mathcal{C}^1$  and  $\mathcal{C}^2$  are shown and the determination of  $\phi(x)$  is illustrated.

Consider the particular case where the plant has a single input  $u^1$ . For each  $N$ ,  $R'_N$  is a convex polyhedron in  $E^n$ . For each  $N = 1, 2, \dots$ ,  $R'_N$  can be obtained from  $R'_{N-1}$  as follows. Translate each vertex of  $R'_{N-1}$  by  $\pm r_N^1$ , the convex hull of this new set of points is  $R'_N$ . In other words, some faces of  $R'_N$  can be obtained from those of  $R'_{N-1}$  by translating them in the direction  $\pm r_N^1$  where the sign of  $\zeta_N^1$  is selected in such a way that the translation occurs in an outward direction with respect to  $R'_{N-1}$ . If, in addition, the single input PAM system is such that  $A$  [in (5)] has real distinct eigenvalues, then the hypersurface  $\mathcal{C}^1$  has special properties that lead to a simple analogue computer set-up to evaluate  $\phi(x)$ <sup>25</sup>. This has been tested experimentally<sup>11</sup> and it was found that the system was not particularly sensitive to changes in the parameters.

Going back to the  $m$  input plant, in order to see what happens to the optimal strategy as the sampling period  $T \rightarrow 0$ , a description is now given of how  $R'_N$  is obtained from  $R'_{N-1}$ . For notational convenience, let  $R'_{N-1} = R'_{N,0}$  and  $R'_N = R'_{N,m}$ . Construct iteratively the sequence of sets  $\{R'_{N,j}, j=1, \dots, m\}$ . Since  $R'_{N,0}$  is a polyhedron in  $E^n$ , it will be seen that each one of the sets of the sequence is a polyhedron.  $R'_{N,j-1}$  is obtained from  $R'_{N,j}$  as follows. Translate each vertex of  $R'_{N,j-1}$  by  $\pm r_{N,j}^j$ , the convex hull of this new set of points is the polyhedron  $R'_{N,j}$ . Performing this operation  $m$  times,  $R'_{N,m} = R'_N$  is obtained from  $R'_{N,0} = R'_{N-1}$ . Now, let  $x \in R'_{N,\gamma}$  but  $x \notin R'_{N,\gamma-1}$ . The scalar  $\beta_N^j$  is uniquely defined

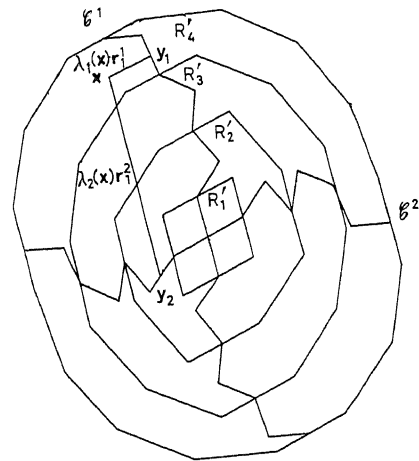


Figure 2. This figure shows the two dimensional state space of a two input system. Note the sets  $R'_1, R'_2, \dots$  and the critical curves  $\mathcal{C}^1$  and  $\mathcal{C}^2$ .  $\lambda_1(x)$  and  $\lambda_2(x)$  are the 'distances' of the state  $x$

\* Throughout the Section 2 of the paper, we shall always write  $r_k^j$  for  $r_k^j(T)$ .

by the following condition\*.  $\beta_N^j$  is the smallest in absolute value of  $\nu$  such that  $x = \nu r_N^j + y$  where  $y \in \partial R_{N-1}^j$ . By the preceding construction,  $|\beta_N^j| \leq 1$ . Proceeding in this manner until one reaches  $R_0^j = \{0\}$ , one gets

$$x = \sum_{k=1}^{N-1} \sum_{j=1}^m \beta_k^j r_k^j + \sum_{j=1}^m \beta_N^j r_N^j \quad (9)$$

where  $\beta_k^j$ 's are uniquely determined and belong to  $[-1, 1]$ . This unique representation of  $x$  is called the canonical representation of  $x$ . It is of interest to note that  $\phi(x)$ , defined earlier in this paper has  $\beta_1^1, \beta_1^2, \dots, \beta_1^m$  as components. An important property of that representation is that, for each  $j$ , the sequence  $\{\beta_k^j\}_{k=1}^N$  has, at most,  $n$  element that are in  $(-1, 1)$ , and all others are either 1 or  $-1$ . Since a PAM system is being considered it means that  $u_{[0, NT]}^j$  differs from  $+1$  or  $-1$  at most over a set of measure  $nT$ . Therefore as  $T \rightarrow 0$ ,  $u_{[0, NT]}^j$  will be bang-bang. In other words, almost everywhere in the state space  $\phi(x)$  has in the limit, components equal to 1 in absolute value.

### Equivalent Optimal Strategies

Let us now go back to some general considerations and examine the concept of an optimal strategy. Let  $f$  be an optimal strategy for the system under consideration. For each  $x$ , the sequence  $f(x)$  is a sequence of vectors  $\{\beta_k\}_{k=1}^N$  (where  $N$  depends on  $x$ ) such that if, at  $(k-1)T$ , ( $k=1, 2, \dots, N$ ), the input to the plant is the modulator output defined by  $\beta_k$  then  $x(NT) = 0$  and no admissible control exists which transfers  $x_0$  to  $0$  in less than  $N$  sampling periods.

In view of the uniqueness theorem of the theory of differential equations and of the assumptions made on  $F(x, u)$  in (1),  $x_1 \neq x_2$  implies that  $f(x_1) \neq f(x_2)$ . Therefore  $f$  is a one-to-one mapping of the state space  $E^n$  onto  $f(E^n)$ . Therefore the inverse map,  $f^{-1}$ , exists:  $f^{-1}$  maps  $f(E^n)$  onto the state space  $E^n$ . This mapping is one-to-one since the system is completely controllable. Note that  $f(E^n)$  is the set of all optimal sequences associated with the optimal strategy  $f$ .

To illustrate these ideas, consider the PAM system and the state  $x_0$  given by eqn (8). The  $mN$  numbers  $\zeta_k^j \alpha_k^j$  define a sequence  $s_N$  of  $N$  elements each one of which is an  $m$ -vector. Thus the right-hand side of (8) associates with each such sequence  $s_N$  a state  $x$ . In other words, by a straightforward analysis of the system, a function  $g$  is obtained, mapping the sequence space  $S$  onto the state space  $E^n$ ,

$$g(s_N) = \sum_{k=1}^N \sum_{j=1}^m \zeta_k^j \alpha_k^j r_k^j$$

Introducing a definition,  $\{\beta_k\}_{k=1}^N$  is said to be a minimal sequence if this sequence transfers some state to  $0$ , and if no other sequence can do it in less than  $N$  sampling periods.

The function  $g$  can be used to construct an optimal strategy as follows. First the domain of  $g$  must be restricted to the set of all minimal sequences. Then, from the properties of  $F$  in (1), the inverse images of all  $x \in E^n$  are disjoint sets, so that if, from each inverse image one and only one minimal sequence is picked arbitrarily, the resulting restriction of  $g$ , say the function  $\hat{g}$ , is one-to-one and finally,  $\hat{g}^{-1}$  is an optimal strategy.

\*  $\beta_k^j$  is defined as the  $j$ th component of the  $\beta_k$  defined earlier in this paper, i.e.,  $\beta_k^j = \zeta_k^j \alpha_k^j$ .

Let  $x_0$  be any initial state, and let  $f(x_0) = \{\beta_k\}_{k=1}^N$  be the sequence assigned to  $x_0$  by the optimal strategy. Let  $\tilde{u}(f(x_0))$  be the control defined by the sequence  $f(x_0)$ . If the system is now steered by the control  $\tilde{u}[f(x_0)]$ , it is clear that the sequences  $\{\beta_k\}_{k=2}^N, \{\beta_k\}_{k=3}^N, \dots, \{\beta_N\}$ , will be minimal for the states  $x_1, x_2, \dots, x_{N-1}$ , and that  $x_N = 0$ . Since a given state may have more than one minimal sequence, it is convenient to assume that the optimal strategy assigns precisely these sequences to these states and that this is true for all  $x_0 \in E^n$ . An optimal-strategy which satisfies this condition is said to be consistent.

**Definition**—Consider two systems  $A$  and  $B$ , with state spaces  $E_A$  and  $E_B$ , with state vectors  $x$  and  $y$ , with consistent optimal strategies  $f_A$  and  $f_B$  and inverse functions  $f_A^{-1}$  and  $f_B^{-1}$ . The systems  $A$  and  $B$  are said to be optimal-strategy-equivalent if  $f_A(E_A) = f_B(E_B)$ .

**Discussion**—Consider any  $x_0 \in E_A$ , since  $f_A(E_A) = f_B(E_B)$ , there exists a  $y_0 \in E_B$  such that  $f_B(y_k) = f_A(x_k)$ ,  $k=0, 1, \dots, N$ , where  $N$  is the length of  $f_A(x_0)$ . Thus, if the transients under the optimal controls  $\tilde{u}[f_A(x_0)]$  and  $\tilde{u}[f_B(y_0)]$  are observed in the sequence space, systems  $A$  and  $B$  are indistinguishable.

Now apply this concept of equivalence to derive an optimal strategy for a system  $B$  from the known optimal strategy of a system  $A$ . Consider a system  $A$  with an optimal strategy  $f_A$  and its inverse  $f_A^{-1}$ . Let  $B$  be any other system, in the class considered, for which it is necessary to construct an optimal strategy, and let  $g_B$  be the function mapping the sequence space  $S$  on to  $E_B$ , such that if  $s_N$  transfers the state  $y$  to  $0$  in  $N$  sampling periods, then  $g_B(s_N) = y$ , (see Figure 3).

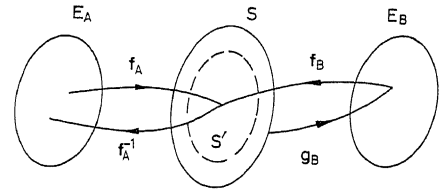


Figure 3. Illustration of the mappings necessary for the definition of equivalence. The set  $S'$  is the set of all optimal sequences under the optimal strategy  $f_A$ .

Clearly, the state space  $E_A$  can be mapped into the state space  $E_B$  by means of the function  $g_B f_A$ . Now let the state of  $A$  be  $x$  and the state of  $B$  be  $y$ . If

$$g_B f_A(E_A) = E_B \quad (10)$$

$$g_B f_A(x_1) \neq g_B f_A(x_2) \text{ for all } x_1, x_2 \text{ such that } x_1 \neq x_2 \quad (11)$$

then the function  $g_B f_A$  is a one-to-one mapping of  $E_A$  on to  $E_B$ . Therefore its inverse exists and one can construct a function  $f_B$  on the state space of  $B$ , such that for all  $x \in E_A$ :

$$f_B(y) = f_A(x) \text{ where } y = g_B f_A(x) \quad (12)$$

Since  $f_A$  is optimal, its inverse  $f_A^{-1}$  exists and, from (12),  $f_A^{-1} f_B$  is the inverse of  $g_B f_A$  and it maps  $E_B$  on to  $E_A$ .

**Theorem 2**—Let  $R_N^A$  be the set  $R_N$  for system  $A$  and let  $R_N^B$  be the set  $R_N$  for system  $B$ , where  $R_N$  was defined earlier in the paper. If for all  $N = 1, 2, \dots$ ,  $g_B f_A(R_N^A) = R_N^B$ , then the function  $f_B$  defined by (12) is an optimal strategy and system  $A$  and  $B$  are equivalent under the optimal strategies  $f_A$  and  $f_B$ .

The usefulness of this theorem lies in the fact that the function  $g_B$  is usually easy to determine, hence once the assumptions of the theorem are satisfied it is easy to derive an optimal strategy for a system from the known optimal strategy of another system.

### An Optimal Strategy for a PWM System with a Linear Plant

In this section single input PWM systems are considered, for which (2) reduces to the form

$$\dot{x} = Ax + e_1 u_1^w(t) \quad (13)$$

$A$  is a  $n \times n$  matrix with non-positive real distinct eigenvalues. In this case,  $s_N = \{\zeta_k^1 \alpha_k^1\}_{k=1}^N$  is a sequence of scalars and the function  $g_w(s_N) = \sum_{k=1}^N \zeta_k^1 r_k^1(\alpha_k^1 T)$ . It is extremely difficult to construct independently an optimal strategy for this system and hence, the optimal strategy  $f_w$  is found from the optimal strategy  $f_a$  for the system,

$$\dot{x} = Ax + e_1 u_1^a(t) \quad (14)$$

where  $A$  is the same as in (5).

Consider the optimal strategy  $f_a$ . In the process of constructing the sets  $R_N^a$ ,  $N = 1, 2, \dots$ , for the PAM systems, it was found that minimal sequences are constructed for all the points in the state space in a unique manner. To clarify this statement, consider any point  $x_0$  in  $R_N^a$ ,  $N = 1, 2, \dots$ . Let the vectors  $\zeta_N^1 \alpha_N^1 r_N^1(T)$ ,  $\zeta_N^1 = \pm 1$ ,  $0 < \alpha_N^1 \leq 1$ , be added to the boundary of  $R_{N-1}^a$  in a scanning motion until it intercepts the given point  $x_0$ . Now let the vector  $\zeta_{N-1}^1 \alpha_{N-1}^1 r_{N-1}^1(T)$ ,  $\zeta_{N-1}^1 = \pm 1$ ,  $0 < \alpha_{N-1}^1 \leq 1$  be added to the boundary  $R_{N-2}^a$  in a scanning motion until it intercepts the point of contact of the vector  $\zeta_N^1 \alpha_N^1 r_N^1(T)$  with the boundary of  $R_{N-1}^a$ , and so forth, using vectors  $\zeta_k^1 \alpha_k^1 r_k^1(T)$ ,  $k = N-2, N-3, \dots, 1$ ,  $\zeta_k^1 = \pm 1$ ,  $0 < \alpha_k^1 \leq 1$ , which are added to the boundaries of the sets  $R_{k-1}^a$ . This process generates a sequence of length  $N$  for the representation of the state  $x_0$ , and, since the state  $x_0$  is by hypothesis in  $R_N^a$ , this sequence is a minimal one and thus the procedure defines an optimal strategy.

In practice it is necessary to compute only the first member of the sequence  $f_a(x_0)$  for any given state  $x_0$  in order to ensure minimal time control. This is much easier to compute than the whole sequence  $f_a(x_0)$  and may be done by adding the vector  $-\zeta_1^1 \lambda_1^1 r_1^1(T)$ ,  $\zeta_1^1 = \pm 1$  to the point  $x_0$ , where  $\lambda_1 > 0$  is a scalar, and by adjusting  $\lambda_1$  until the vector  $[x_0 - \zeta_1^1 \lambda_1^1 r_1^1(T)]$  touches the  $\mathcal{C}_a^1$ . If the resulting  $\lambda_1 \geq 1$ ,  $\alpha_1^1 \zeta_1^1 = \zeta_1^1$ , and if the resulting  $\lambda_1 < 1$ , then  $\zeta_1^1 \alpha_1^1 = \zeta_1^1 \lambda_1$ . Thus,  $\alpha_1^1$  is a function of the distance of the point from the  $\mathcal{C}_a^1$ . It is clear that the remainder of the sequence for any given point could be computed by iteration.

Now some useful properties of the vectors  $r_1^1(t)$  are described. For the system (13) it can be shown that

$$r_k^1(t) = r_1^1[t + (k-1)T] - r_1^1[(k-1)T] \quad (15)$$

**Theorem 3**—Let  $c_{ij}$  be a chord on  $r_1^1(t)$ ,  $0 \leq t < \infty$ , such that  $c_{ij} = r_1^1(t_j) - r_1^1(t_i)$ ,  $t_j > t_i$ . Consider any set  $\{c_{i_k j_k}\}_{k=1}^n$  of  $n$  such chords where  $t_{i1} < t_{j1} < t_{i2} < t_{j2} < \dots < t_{jn}$ ; the set  $\{c_{i_k j_k}\}_{k=1}^n$  forms a basis for the space  $E_w$ .

Introducing a mapping of the sets  $R_N^a$  into  $E_w$ , let  $R_N^w$  be the set  $R_N$  in  $E_a$  and let  $R_N^w$  be the set  $R_N$  in  $E_w$ . Let the parametrized arc traced out by the vector  $\zeta_k^1 r_k^1(\alpha_k^1 T)$  for  $0 \leq \alpha_k^1 \leq 1$  be denoted by  $\zeta_k^1 L_k^1$ ,  $k = 1, 2, \dots$ ,  $\zeta_k^1 = \pm 1$ . Consider some

sequence  $s_N \in S$ , then, to map this sequence into  $E_a$ , the sum  $\sum_{k=1}^N \zeta_k^1 \alpha_k^1 r_k^1(T)$  is formed, i.e., geometrically, the parametrized straight line segments  $r_k^1(T)$ ,  $k = 1, 2, \dots, N$ , are put together in the manner prescribed by the sum. In order to map this same

sequence  $s_N$  into  $E_w$ , the sum  $\sum_{k=1}^N \zeta_k^1 r_k^1(\alpha_k^1 T)$  is formed, or, again speaking geometrically, the parametrized arcs  $\zeta_k^1 L_k^1$ ,  $k = 1, 2, \dots$  are put together in the manner prescribed by the sum. Thus, in order to map the sets  $R_N^a$  into  $E_w$  by means of the function  $g_w f_a$ , the arcs  $L_k$  are put together in  $E_w$  in the same canonical representation as the vectors  $r_k^1(T)$  were put together in  $E_a$  to form the sets  $R_N^a$ . It is clear that there is no ambiguity as to direction, since the vectors  $r_k^1(T)$  form chords on the arcs  $L_k$ .

Consider what happens when the arc  $L_N$ ,  $N = 1, 2, \dots$ , is added to the map of one face of the convex polyhedron  $U R_N^a$ . The face of  $U R_N^a$  in  $E_a$  is at most an  $(n-1)$ -dimensional polyhedron and its map in  $E_w$  is formed by a combination of at most  $(n-1)$  of the arcs  $\zeta_k^1 L_k^1$ ,  $\zeta_k^1 = \pm 1$ ,  $k = 1, 2, \dots, N-1$ . Hence, when the arc  $\zeta_N^1 L_N^1$ ,  $\zeta_N^1 = \pm 1$  is added to any point of the map of the face of  $R_{N-1}^a$ , it follows from Theorem 3 that it does not intersect this map at any other point. It also follows from Theorem 3 that if this addition of the arc  $\zeta_N^1 L_N^1$  is done in a continuous scanning motion, a hypervolume is traced out, each point of which is intercepted only once by the arc  $\zeta_N^1 L_N^1$ . Since the vertices of the map of  $R_N^a$  are the same as the vertices of  $R_{N-1}^a$ , if added outwards to one face of the map of  $R_{N-1}^a$ , the arc  $L_N$  will not intersect any other face of the map of  $R_{N-1}^a$ . Furthermore, it is clear that if the mapping process is continued, the maps of the sets  $R_N^a$  will fill the space  $E_w$ . Hence it follows that  $g_w f_a(E_a) = E_w$  and that given any  $x_1, x_2$  in  $E_a$ ,  $g_w f_a(x_1) \neq g_w f_a(x_2)$  if  $x_1 \neq x_2$ . It is therefore possible to define a function  $f_w$  on  $E_w$  such that

$$f_w(g_w f_a(x_0)) = f_a(x_0), x_0 \in E_a \quad (16)$$

This is the function  $f_B$  defined in eqn (12) for the general case. Now, it will be observed that as the arcs  $\zeta_k^1 L_k^1$ ,  $\zeta_k^1 = \pm 1$  are put together to form the maps of the sets  $R_N^a$ , they produce hypervolumes which contain all points expressible in the form  $\sum_{k=1}^N \zeta_k^1 r_k^1(\alpha_k^1 T)$ ,  $\zeta_k^1 = \pm 1$ ,  $0 \leq \alpha_k^1 \leq 1$ . It follows that the maps of the sets  $R_N^a$  are the sets  $R_N^w$  and hence the sets  $R_N$  are invariant under the map  $g_w f_a$ . It may therefore be concluded that the function  $f_w$  defined in (16) is indeed an optimal strategy for the PWM system.

Consider now the distance function required for the computation of  $\zeta_1^1 \alpha_1^1$  for the PWM system. The map of  $E_a$  on to  $E_w$  by means of the function  $f_w f_a$  is clearly, one-to-one and bicontinuous. Hence, the map of the critical hypersurface  $(\mathcal{C}_a^1)$  from  $E_a$  into  $E_w$  results in a critical hypersurface  $(\mathcal{C}_w^1)$  in  $E_w$  with the same properties with respect to  $f_w$  as  $\mathcal{C}_a^1$  has with respect to  $f_a$ . Furthermore, it is clear that all the states  $x$  of  $E_w$ , for which  $f_w(x) = \{\zeta_k^1 \alpha_k^1\}_{k=1}^N$ ,  $N = 1, 2, \dots$ , is such that  $0 \leq \alpha_k^1 < 1$ , are assigned by  $f_w$  a representation of the form  $x = \zeta_1^1 r_1^1(\alpha_1^1 T) + y_1$ , where  $y_1$  is some vector such that  $y_1 \in \mathcal{C}_w^1$  and  $\zeta_1^1 = \pm 1$ ,  $0 \leq \alpha_1^1 < 1$ .

It therefore follows that in order to compute the first term  $\zeta_1^1 \alpha_1^1$  of the sequence  $f_w(x_0)$  for the state  $x_0$ , one forms the sum  $[x_0 - \zeta_1^1 r_1^1(t)]$ ,  $0 \leq t \leq \infty$ , and one finds the value of  $t$  for which

$[x_0 - \zeta_1^1 r_1^1(t)] \in \mathcal{C}_w^1$ . Then, if  $t \geq 1$ ,  $\zeta_1^1 \alpha_1^1 = \zeta_1^1$ , and if  $t < 1$ ,  $\zeta_1^1 \alpha_1^1 = \zeta_1^1 t$ . Again, it is clear that the remainder of the minimal sequence for the point  $x_0$  could be computed by an iterative technique.

### A Few Extensions

Consider the problem of transferring in minimum time an initial state  $x_0$  to a fixed target state  $x_t$  by an admissible control. The set of all admissible controls which transfers  $x_0$  to  $x_t$ , say  $\Delta(x_0; x_t)$ , may be empty. The problem of existence of such admissible controls is the analogue of the controllability problem in the first section of this paper. If the target state  $x_t$  belongs to the set  $\Omega$ , where  $\Omega$  is the set of all states reachable from the origin by admissible controls of a completely controllable system, then  $\Delta(x_0; x_t)$  is not empty for any  $x_0$  in the state space.

Consider a system described by eqn (5). For any  $x_t$ , for which  $\Delta(x_0; x_t)$  is not empty, the problem reduces to the minimal time regulator problem. The procedure is as follows. First, determine the minimum  $N$ , say  $M$ , such that  $(x_0 - A^{-M+1} x_t) \notin R'_{M-1}$  but  $(x_0 - A^{-M} x_t) \in R'_M$ ; secondly, consider the state  $x_0 - A^{-M} x_t$  as the initial state of the system for the minimal time regulator problem; thirdly, the optimal strategy for transferring  $x_0$  to  $x_t$  in  $M$  sampling periods is obtained iteratively as follows. The optimal control for the first sampling period is the optimal control for transferring the state  $x_0 - A^{-M} x_t$  to 0, namely  $\phi(x_0 - A^{-M} x_t)$ .

Note that the resulting sequence is not the optimal sequence associated with transferring  $x_0 - A^{-M} x_t$  to 0.

Earlier in this paper an optimal strategy for a PWM system was derived from that of a PAM system, each system having identical linear plants. However, for two systems to be equivalent, the plant of the first need not be identical with the plant of the other system. In fact<sup>31</sup>, a second-order PWM system with a linear plant, say,

$$\begin{bmatrix} \dot{x}_1 \\ \dot{x}_2 \end{bmatrix} = \begin{bmatrix} 0 & 1 \\ 0 & -a \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} + u^1(t) \begin{bmatrix} 0 \\ 1 \end{bmatrix}$$

is optimal strategy equivalent to the second order PWM system with a non-linear plant

$$\begin{bmatrix} \dot{x}_1 \\ \dot{x}_2 \end{bmatrix} = \begin{bmatrix} x_2 \\ -g(x_2) \end{bmatrix} + u^1(t) \begin{bmatrix} 0 \\ 1 \end{bmatrix}$$

provided  $g(x_2)$  is odd and has a continuous derivative  $g'(x_2) > 0$  for all  $x_2$ .

### References

- <sup>1</sup> WIENER, N. *The Extrapolation, Interpolation and Smoothing of Stationary Time Series with Engineering Applications*. 1949. New York; Wiley
- <sup>2</sup> KOLMOGOROV, A. Interpolation und Extrapolation von Stationären zufälligen Folgen. *Bull. Acad. Sci. U.R.S.S.* 5 (1941) 3-14
- <sup>3</sup> McDONALD, D. C. Nonlinear techniques for improving servo-performance. *Proc. nat. Electron. Conf.* 6 (1950) 400-421
- <sup>4</sup> BELLMAN, R. *Dynamic Programming*. 1957. Princeton; Princeton University Press
- <sup>5</sup> BOLTYANSKII, V. G., GAMKRELIDZE, R. V. and PONTRYAGIN, L. S. Teoria optimal'nikh protsessov. *Izv. Akad. Nauk SSSR. Seria Matematicheskaya*, TOM 24 (1960) 1-42
- <sup>6</sup> LA SALLE, J. P. The bang-bang principle. *RIAS Report*. 1959
- <sup>7</sup> BELLMAN, R., GLICKSBERG, I., and GROSS, O. On the bang-bang control problem. *Quart. J. appl. Math.* 14 (1956) 11-18
- <sup>8</sup> DESOER, C. A. The bang-bang control problem treated by variational techniques. *Information and Control* 2 (1959) 333-348
- <sup>9</sup> KALMAN, R. E. On the general theory of control systems. *Automatic and Remote Control*. Vol. I, p. 481. 1961. London; Butterworths
- <sup>10</sup> KALMAN, R. E. Canonical structures of linear dynamical systems. *Proc. nat. Acad. Sci., Wash.* 48 (1962) 596-600
- <sup>11</sup> CHANG, R. L. Analog computer study of an optimal strategy for a sampled data system. *M. S. Thesis*. 1961. Berkeley; University of California
- <sup>12</sup> ATHANASSIADES, M., and SMITH, M. J. O. Theory and design of high order bang-bang control systems. *Trans. Inst. Radio Engrs N. Y. AC-6*, 2 (May 1961) 125-135
- <sup>13</sup> CHANG, S. S. L. Computer optimization of nonlinear control systems by means of digitized maximum principle. *Inst. Radio Engrs Int. Conv. Rec. Pt 4*, p. 48-56
- <sup>14</sup> HO, YU-CHI. A successive approximation technique for optimal control systems subject to input saturation. *Trans. Amer. Soc. mech. Engrs; J. Basic Eng.* 84, Series D, 1 (1962) 33-41
- <sup>15</sup> EATON, J. H., and ZADEH, L. A. Optimal pursuit strategies in discrete-state probabilistic systems. *Trans. Amer. Soc. mech. Engrs; J. Basic Eng.* 84, Series D, 1 (1962) 23-29
- <sup>16</sup> BREAKWELL, J. V. The optimization of trajectories. *J. Soc. Ind. appl. Math.* 7 (1959), 215-247
- <sup>17</sup> BRYSON, A. E., and DENHAM, W. F. *A Steepest Ascent Method for Solving Optimum Programming Problems* (to be published)
- <sup>18</sup> KRASOVSKII, N. N. Ob Odnoi Zadache Optimal'noy Regulirovaniya. *Prikl. Math. i Mekh.* 21, 5 (1957) 670-677
- <sup>19</sup> KRASOVSKII, N. N. On the theory of optimum control. *Prikl. Math. i Mekh.* 23, 4 (1959) 625-639
- <sup>20</sup> TSYPKIN, J. Z. Optimal'nye Protsesy v Impul'snikh Avtomaticheskikh Sistemakh. *Izv. Akad. Nauk SSSR, OTN Energetika i Avtomatika* 4 (1960)
- <sup>21</sup> MOORE, E. F. Shortest path through a maze. *Proc. Int. Symp. on Switching Circuits*. Harvard University, April 1957
- <sup>22</sup> KALMAN, R. E. Optimum nonlinear control of saturating systems by intermittent action. *Inst. Radio Engrs Wescon Convention Record*, 1, Pt 4 (1957) 130-135
- <sup>23</sup> NELSON, W. L. Optimal control methods for on-off sampling systems. *Trans. Amer. Soc. mech. Engrs; J. Basic Eng.* 84, Series D, 1 (1962) 91-101
- <sup>24</sup> DESOER, C. A., and WING, J. An optimal strategy for a saturating sampled-data system. *Trans. Inst. Radio Engrs PGAC* 6, 1 (1961) 5-15
- <sup>25</sup> DESOER, C. A., and WING, J. A minimal time discrete system. *Inst. Radio Engrs, Trans. Aut. Contr.* AC-6, 2 (1961) 111-125
- <sup>26</sup> DESOER, C. A., and WING, J. The minimal time regulator problem for linear sampled data systems: General theory. *J. Franklin Inst.* 272 (1961) 208-228
- <sup>27</sup> WING, J., and DESOER, C. A. The Multiple Input Minimal Time Regulator Problem: General Theory I.E.E.E. *Trans. Aut. Contr.* AC-8, No. 2 (1963) 125-136
- <sup>28</sup> POLAK, E. Minimum time control of second order pulse-width-modulated sampled-data systems. *Trans. Amer. Soc. mech. Engrs; J. Basic Eng.* 84, Series D (1962) 101-110
- <sup>29</sup> POLAK, E. On the equivalence of discrete systems in time-optimal control. *Trans. Amer. Soc. mech. Engrs J. Basic Eng.* Series D, (1963) 208-210
- <sup>30</sup> CODDINGTON, E. A., and LEVINSON, N. *Theory of Ordinary Differential Equations*. 1955. New York; McGraw-Hill
- <sup>31</sup> POLAK, E. Optimal time control of some pulse width modulated sampled-data systems. *Ph. D. Thesis* (Electrical Engineering) 1961. Berkeley; University of California Press
- <sup>32</sup> POLAK, E. *Minimal Time Control of a Discrete System with a Non-linear Plant* I.E.E.E. *Trans. Aut. Contr.* AC-8, No. 1 (1963) 49-56

## DISCUSSION

YA. Z. TSYPKIN, *Institute of Automatics and Telemechanics, Moscow I-15, Kalantskovskaja 15a, U.S.S.R.*

I should like to ask the following questions:

- (1) What is the nature of PAM and PWM optimal processes?
- (2) Did you compare your proposed optimal systems with similar systems using linear control laws?
- (3) What is the nature of the processes in such an optimal system when it is not excited by the presence of an initial condition, but by means of an external excitation such as, for example, a monotonic input?

C. A. DESOER and E. POLAK, *in reply*

We thank Professor Tsyppkin for his interest and his stimulating questions, which we answer as follows:

First, the control signal generated by our proposed optimal strategy has the following properties: (1) not more than  $n$  of the  $\alpha_k$ 's are less than 1, and all other  $\alpha_k$ 's are equal to 1; (2) for real eigenvalues, the sequence  $\{\xi_k\}$  has, at most,  $n-1$  sign reversals; (3) for complex eigenvalues, given any number, say  $M$ , initial states can be found such that the associated control signal will have more than  $M$  sign reversals.

Second, we have not obtained analytical results which allow us to compare conveniently our optimal systems with similar systems governed by linear control laws. However, our algorithm to construct the sets  $R_N$  can be used to compile a control table for an arbitrary bounded portion of the state space. This table may then be used to obtain an average cost for the system. This figure of merit may then be used for comparison with other control laws.

Third, a second-order PAM system has been simulated on an analogue computer. Its behaviour with respect to standard inputs, noise inputs, and parameter perturbations were found to be satisfactory from an engineering point of view. Our experimental investigations are at present being expanded to include PWM systems and systems of higher order. We hope to be in a position to report on our experimental results in the near future.

R. G. MARIOUW, *Ministry of Defence, Juliana van Stolberglaan 381, Den Haag, Netherlands*

In their extremely interesting paper on minimum time discrete regulators, Professor Desoer and his co-workers consider the optimization of a system described by the equation

$$\dot{x} = Bx + Eu(t)$$

including a modulator of either the P.A.M. or P.W.M. type.

However, an important practical case obtained, for instance, in a controlled rectifier modulator, is that of pulse phase / amplitude modulation (PPAM) in which there is a time delay of the form

$$H\{t - (K|u| + a)T\}$$

associated with the control vector  $u(t)$ , whilst furthermore there is a constraint on the maximum modulus of  $u(t)$ . In my opinion, the control matrix  $E$  will then be no longer constant.

Can the authors please comment on this question and possibly show the way in which an answer might be obtained?

C. A. DESOER and E. POLAK, *in reply*

We thank Mr. Mariouw for his discussion and for bringing to our attention this practically motivated modulation scheme. Unfortunately, his description does not define precisely the modulation scheme involved. However, since controlled rectifiers are mentioned, the modulation scheme might be such that the pulse width is proportional to the input and the leading edge is delayed proportionately to the input. In our analysis of PWM systems, the leading edge was fixed and the lagging edge was delayed proportionately to the input. Because of the similarity of the resulting non-linear difference equations, it seems, therefore, that some of the techniques of our paper should apply to this problem.



# On the Roots of a Real Polynomial Inside the Unit Circle and a Stability Criterion for Linear Discrete Systems\*

E. I. JURY

## Summary

Necessary and sufficient algebraic conditions for the roots of a real polynomial to lie inside the unit circle are given in table form. In this form, the constraints are obtained only by evaluation of second-order determinants. The connection and identity existing between the stability criterion established in this paper and that of a previously obtained criterion<sup>1</sup> are discussed. The usefulness of the table may be found in those cases where the coefficients of the real polynomial are given in numbers. It is similar to Routh's table obtained for the continuous case.

Conditions on the numbers of the roots inside, outside, or on the unit circle are also discussed, within the cases when the determinants are zero or non-zero. Also, necessary and sufficient conditions are formulated for all the roots to be inside a circle of radius  $\sigma$  less than unity, and also the conditions when the roots are to lie between plus and minus unity in the  $z$  plane.

Various examples of discrete systems are presented which illustrate the applications of the new stability criterion as well as the other conditions formulated in this paper.

In concluding the paper, various analytical stability criteria applied to linear discrete systems are enumerated and compared, with emphasis on the advantageous applications of each.

## Sommaire

Les conditions algébriques nécessaires et suffisantes pour que les racines d'un polynôme réel se trouvent à l'intérieur du cercle unitaire, sont données sous la forme d'un tableau. Sous cette forme les limitations sont obtenues par l'évaluation d'un déterminant du 2<sup>e</sup> ordre seulement. La relation existant entre le critère de stabilité décrit dans ce rapport et un critère déjà connu, est examinée. Ce tableau est utile dans les cas où les coefficients du polynôme à analyser sont connus sous forme numérique. Il rejoint le critère de stabilité de Routh dans le cas continu.

Le nombre des racines situées à l'intérieur, à l'extérieur et sur le cercle unitaire, est discuté et cela aussi bien dans les cas où les déterminants sont nuls, que lorsqu'ils sont différents de zéro. En outre, les conditions nécessaires et suffisantes sont formulées pour que toutes les racines soient à l'intérieur d'un cercle de rayon  $\sigma$  inférieur à l'unité; il est également indiqué quelles sont les conditions pour que les racines soient comprises entre plus ou moins un dans le plan  $z$ .

Différents exemples de systèmes discrets sont indiqués pour illustrer l'application de ce nouveau critère de stabilité et les autres conditions décrites dans ce rapport.

En conclusion, différents critères de stabilité analytiques applicables à des systèmes linéaires discrets sont énumérés et comparés en mettant en évidence les avantages et les inconvénients de l'application de chacun d'eux.

## Zusammenfassung

Die notwendigen und hinreichenden algebraischen Bedingungen dafür, daß die Wurzeln eines reellen Polynoms innerhalb des Einheitskreises liegen, werden in Tabellenform angegeben. Zur Ermittlung der Stabilitätsbedingungen sind dabei nur Determinanten zweiter Ordnung zu lösen. Die Zusammenhänge und die Entsprechung zwischen dem hier aufgestellten Stabilitätskriterium und einem früher erhaltenen Kriterium<sup>1</sup> werden dargestellt. Diese Tabelle mag in solchen Fällen nützlich

sein, in denen die Koeffizienten des reellen Polynoms zahlenmäßig vorliegen. Sie ähnelt der Routhschen Tabelle für den kontinuierlich arbeitenden Fall.

Bedingungen für die Anzahl der Wurzeln innerhalb, außerhalb und auf dem Einheitskreis werden für die Fälle, daß die Determinanten Null oder von Null verschieden sind, besprochen. Es werden auch die notwendigen und hinreichenden Bedingungen dafür formuliert, daß sich alle Wurzeln innerhalb eines Kreises mit dem Radius  $\sigma < 1$  befinden und dafür, daß die Wurzeln in der  $z$ -Ebene zwischen  $+1$  und  $-1$  liegen.

An Hand von verschiedenen Beispielen diskreter Systeme werden die Anwendung dieses neuen Stabilitätskriteriums wie auch der anderen hier formulierten Bedingungen gezeigt.

Abschließend werden verschiedene analytische Stabilitätskriterien für lineare diskrete Systeme aufgezählt und mit besonderem Hinweis auf die jeweils günstigsten Anwendungsfälle miteinander verglichen.

## Introduction

A useful criterion to obtain the number of roots of a real polynomial in the unit circle has been long recognized and desired. The early work of Schur-Cohn<sup>2</sup>, Fujiwara<sup>3</sup>, Marden<sup>4</sup> and others is an indication of the importance of such a criterion. Since heretofore no simple criteria had been obtained, many authors in the past have used transformation to a different plane, in which several criteria were either available or readily obtained. Among the general transformations, the bi-linear and the unit shifting are sometimes used. The advantage of the bi-linear<sup>5</sup> transformation, which maps the inside of the unit circle into the left half of the  $W$  plane, is the ready use of the Hurwitz, Liénard-Chipart, or Routh tables. However, this bi-linear transformation involves complicated algebraic manipulations which increase in complexity as the order of the real polynomial becomes higher. The stability constraints within the shifted unit circle, although useful, require the formulation of sampled-data system in the shifted  $z$  plane<sup>6, 7</sup>.

In a recent work on the problem of stability of linear discrete systems, or alternately, the necessary and sufficient conditions for the roots of a real polynomial to be inside the unit circle, the author has succeeded in obtaining a satisfactory criterion<sup>1</sup>. This criterion constitutes a major simplification of the Schur-Cohn criterion, and yields the final constraints in a usable form such that no transformation is required. It is similar to the Hurwitz or Liénard-Chipart criteria<sup>8</sup> and has the advantage that it may be readily used for the design of linear discrete systems.

\* A portion of this paper has been obtained from the contents of the report, 'On the Roots of a Real Polynomial Inside the Unit Circle and a Stability Criterion for Linear Discrete Systems', by E. I. Jury and J. Blanchard, ERL Report No. 425, Issue No. 60, University of Calif., Berkeley; December 26, 1961.



In this paper an alternate criterion for stability or for obtaining the number of roots inside the unit circle is obtained. This criterion is similar to the Routh table used for the continuous case in obtaining the number of roots in the left half of the  $s$  plane. It is given in table form and can be calculated by expanding only second-order determinants. It has the advantage over the preceding criterion in the sense that calculation of higher order determinants is no longer required. Furthermore, it is particularly useful if the coefficients of the polynomial are given in numbers.

The derivation of the criterion is based on the theorems of Rouché, which were discussed extensively by Marden<sup>4</sup>. The latter has obtained a table form of the criterion for the complex coefficients and Tsytkin<sup>9</sup> had rearranged this table for the real coefficient case. Recently Brown<sup>10</sup> has applied the Marden table to the stability test of a fourth order real polynomial. This author has obtained a major simplification of this earlier work and has shown the connection between the table form and the earlier obtained simplified criterion. The application of the criterion in the design of linear discrete systems as well as other applications constitutes the major discussion of this paper.

### Mathematical Background

The work of Marden<sup>4</sup> which is based on the earlier work of Rouché<sup>11</sup> is first reviewed. Only real polynomials will be considered in this paper; however, some of the theorems can also be applied to the general coefficient case. Let the considered polynomial be denoted as  $F(z)$  and be written as follows:

$$F(z) = a_0 + a_1 z + a_2 z^2 + \dots + a_k z^k + \dots + a_n z^n$$

Consider the sequence of polynomials  $F_0(z), F_1(z), \dots, F_j(z)$ . If to each of these polynomials is associated the polynomial  $F_j^*(z)$ , whose roots are the reciprocal ones of  $F_j(z)$ , then  $F_{j+1}(z)$  can be obtained from  $F_j(z)$  as follows:

$$F_{j+1}(z) = a_0^{(j)} F_j(z) - a_n^{(j)} F_j^*(z), \quad j=0, 1, \dots, n-1 \quad (1)$$

where

$$F_j(z) = \sum_{k=0}^{n-j} a_k^{(j)} z^k \quad (1a)$$

It should be noted that all  $a_k^{(0)}$  are written as  $a_k$ .

$$F_j^*(z) = z^{n-j} F_j(1/z), \quad j=0, 1, \dots, n-1 \quad (2)$$

The following observations, which become evident if, for example,  $F_j(z)$  is divided by  $F_j^*(z)$  for  $j=0$  as defined in (2), can be deduced from eqns (1) and (2):

(a) If  $F_j(z)$  is of degree  $n-j$ , then  $F_{j+1}$  is at most of degree  $n-j-1$ .

(b) The coefficients of  $F_{j+1}$  can be obtained from the coefficients of  $F_j$  as follows:

$$a_k^{(j+1)} = a_0^{(j)} a_k^{(j)} - a_n^{(j)} a_{n-j-k}^{(j)} \quad (3)$$

(c)  $F(z)$  is equivalent to  $F_0(z)$ , by the definition of  $F_j(z)$  from (1a).

In each polynomial  $F_j(z)$ , the constant real coefficient  $a_0^{(j)}$  is denoted by  $\delta_j$ , such that

$$\delta_{j+1} = [a_0^{(j)}]^2 - [a_n^{(j)}]^2 = a_0^{(j+1)} \quad (4)$$

(d) The zeros of  $F_j^*(z)$  are relative to the circle  $|z|=1$ , the inverses of the zeros of  $F_j(z)$ .

The following theorem has been proved by Marden<sup>4</sup>. If for the polynomial  $F(z)$ ,  $p$  of the products  $(P_k = \delta_1 \delta_2 \dots \delta_k; k=1, 2, \dots, n)$  are negative and the remaining  $n-p$  are positive, then  $F(z)$  has  $p$  zeros in the unit circle  $|z|=1$ , no zeros on this circle and  $n-p$  zeros outside this circle. From this theorem it can be deduced that the necessary and sufficient condition for  $F(z) = 0$  to have all its roots inside the unit circle is

$$\delta_1 < 0, \delta_2 > 0, \dots, \delta_n > 0 \quad (5)$$

A convenient way to obtain  $\delta_k$  and  $F_j(z)$  is naturally to construct the following table by noting eqns (2)-(4).

Table 1. A Stability Procedure Form

Polynomial	Row	$\delta_k$	Coefficients					
$F(z)$	1		$a_0$	$a_1$	$a_2 \dots$	$a_{n-k}$	$\dots$	$a_n$
$F^*(z)$	2		$a_n$	$a_{n-1}$	$\dots$	$a_k$	$\dots$	$a_0$
$F_1(z)$	3	$\delta_1 = a_0^{(1)}$	$a_0^{(1)}$	$a_1^{(1)}$	$\dots$			$a_{n-1}^{(1)}$
$F_1^*(z)$	4		$a_{n-1}^{(1)}$	$a_{n-2}^{(1)}$	$\dots$			$a_0^{(1)}$
.		$\delta_2 = a_0^{(2)}$						
.		$\vdots$						
.		$\vdots$						
$F_{n-1}(z)$	$2n-1$	$\delta_{n-1} = a_0^{(n-1)}$	$a_0^{(n-1)}$	$a_1^{(n-1)}$				
$F_{n-1}^*(z)$	$2n$		$a_1^{(n-1)}$	$a_0^{(n-1)}$				
$F_n(z)$	$2n+1$	$\delta_n = a_0^{(n)}$	$a_0^{(n)}$					
$F_n^*(z)$	$2n+2$		$a_0^{(n)}$					

In Table 1,

$$a_0^{(1)} = \begin{vmatrix} a_0 & a_n \\ a_n & a_0 \end{vmatrix}, \quad a_k^{(1)} = \begin{vmatrix} a_0 & a_{n-k} \\ a_n & a_k \end{vmatrix}, \quad a_0^{(2)} = \begin{vmatrix} a_0^{(1)} & a_{n-1}^{(1)} \\ a_{n-1}^{(1)} & a_0^{(1)} \end{vmatrix}, \quad \dots$$

Furthermore, it should be noted that the elements of row  $2k+2$  consist of the coefficients of the row  $2k+1$  written in reverse order.  $k=0, 1, 2, \dots, n$ .

The number of roots inside the unit circle if  $\delta_k \neq 0$ , is equal to the negative number of the products of

$$P_k = \delta_1 \delta_2 \delta_3, \dots, \delta_k, \quad k=1, 2, 3, 4, \dots, n \quad (6)$$

where  $P_1 = \delta_1, P_2 = \delta_1 \delta_2, \dots$

The necessary and sufficient condition for all the roots to be inside the unit circle is when the negative number of the products  $P_k$  is  $n$ . This condition is satisfied if and only if the following exists:

$$\delta_1 < 0, \delta_2 > 0, \delta_3 > 0, \dots, \delta_n > 0 \quad (7)$$

### Example 1

To illustrate the use of the above table, consider the following polynomial:

$$F(z) = 3 - 2z - \frac{3}{2}z^2 + z^3 \quad (8)$$

Polynomial	Row	k	Coefficients			
$F(z)$	1		3	-2	$-\frac{3}{2}$	1
$F^*(z)$	2		1	$-\frac{3}{2}$	-2	3
$F_1(z)$	3	$\delta_1 = 8$	8	$-\frac{9}{2}$	$-\frac{5}{2}$	
$F_1^*(z)$	4		$-\frac{5}{2}$	$-\frac{9}{2}$	8	
$F_2(z)$	5	$\delta_2 = \frac{231}{4}$	$\frac{231}{4}$	$-\frac{189}{4}$		
$F_2^*(z)$	6		$-\frac{189}{4}$	$\frac{231}{4}$		
$F_3(z)$	7	$\delta_3 \cong 1,100$	$\cong 1,100$			

In this case,  $\delta_1 > 0$ ,  $\delta_2 > 0$ ,  $\delta_3 > 0$  therefore

$$\begin{aligned} P_1 &= \delta_1 > 0 \\ P_2 &= \delta_1 \delta_2 > 0 \\ P_3 &= \delta_1 \delta_2 \delta_3 > 0 \end{aligned} \quad (9)$$

Since all the  $P_k$ 's are positive, therefore none of the roots are inside the unit circle. The roots of the above equation are  $z_1 = 3/2$ ,  $z_{2,3} = \pm \sqrt{2}$ , which verifies that they are all outside the unit circle.

#### Simplification of the Stability Criterion and the Table Form<sup>12</sup>

To simplify the criterion for the number of roots or for the stability, the connection between the  $\delta_k$  given in this paper and Schur-Cohn<sup>2</sup> determinants  $\Delta_k$ 's is established. The determinants  $\Delta_k$ 's are discussed extensively in the literature<sup>1, 4</sup> and they offer an alternate method for obtaining the number of roots inside the unit circle. The connection between  $\delta_k$  and  $\Delta_k$  was first established and proved by Marden<sup>4</sup> and is given by the following relation:

$$\Delta_k = \frac{\delta_k}{\delta_1^{k-2} \delta_2^{k-3} \dots \delta_{k-2}} \quad (10)$$

This equation could be used to calculate the  $\Delta_k$ 's from  $\delta_k$ 's, thus avoiding evaluation of higher order determinants. When  $k = n$ , the above yields

$$\Delta_n = \frac{\delta_n}{\delta_1^{n-2} \delta_2^{n-3} \delta_3^{n-4} \dots \delta_{n-2}} \quad (11)$$

or

$$\delta_n = \delta_1^{n-2} \delta_2^{n-3} \delta_3^{n-4} \delta_4^{n-5} \dots \delta_{n-3}^2 \delta_{n-2} \Delta_n \quad (12)$$

In an earlier work<sup>1</sup> it is proved that

$$\Delta_n = A_n^2 - B_n^2 = (A_{n-1} - B_{n-1})^2 F(1) \cdot F(-1) \quad (13)^*$$

Substituting this expression in (12), the following is obtained

\* The stability constants  $A_n, B_n, A_{n-1}, B_{n-1}$  are defined and discussed in earlier investigations<sup>1</sup>.

$$\delta_n = \delta_1^{n-2} \delta_2^{n-3} \delta_3^{n-4} \dots,$$

$$\delta_{n-2} (A_{n-1} - B_{n-1})^2 F(1) \cdot F(-1), \quad n \geq 2 \quad (14)$$

Since  $(A_{n-1} - B_{n-1})^2$  is always positive, and for the roots to be inside the unit circle,  $\delta_2, \delta_3, \dots, \delta_{n-1}$  are all positive; therefore, the last condition  $\delta_n > 0$  can be equivalently replaced by:

$$\delta_n > 0 \Leftrightarrow (-1)^n F(1) \cdot F(-1) > 0 \quad (15)$$

If we let  $a_n > 0$ , then  $\delta_n > 0$  reduces to<sup>15</sup>

$$F(1) > 0, F(-1) > 0, n \text{ even} \\ < 0, n \text{ odd} \quad (16)$$

It should be noted that the last constraint in eqn (7) could be replaced by the above equation by letting  $a_n > 0$ .

The above simplification may also be utilized to obtain information on the number  $P_n$  without having to calculate  $\delta_n$ , but using instead relationship (14).

Since the connection between  $\delta_k$  and  $\Delta_k = A_k^2 - B_k^2$  is obtained, one can apply all the significant simplifications obtained<sup>1</sup> on the  $A_k$ 's and  $B_k$ 's\* to  $\delta_k$  or to the stability problem using this criterion. This connection will be shown in a few examples.

Utilizing eqn (16), one can rewrite Table 1 in the form of Table 2 by noting the following from Tables 1 and 2:

$$b_k = a_k^{(1)}, c_k = a_k^{(2)}, d_k = a_k^{(3)}, e_k = a_k^{(4)}, \dots$$

and

$$\delta_1 = a_0^{(1)} = b_0 = a_0^2 - a_n^2, \delta_2 = a_0^{(2)} = b_0^2 - b_{n-1}^2,$$

$$\delta_3 = a_0^{(3)} = c_0^2 - c_{n-2}^2, \dots$$

$$\delta_{n-1} = a_0^{(n-1)} = r_0^2 - r_2^2,$$

$$\delta_n = (r_0^2 - r_2^2)^2 - (r_0 r_1 - r_1 r_2)^2 = a_0^{(n)}$$

The necessary and sufficient conditions for the roots of  $F(z) = 0$ , with  $a_n > 0$ , to lie inside the unit circle are given as follows:

$$F(1) > 0, F(-1) > 0, n \text{ even} \\ < 0, n \text{ odd} \quad (17)$$

$$\left. \begin{aligned} |a_0| &< a_n \\ |b_0| &> |b_{n-1}| \\ |c_0| &> |c_{n-2}| \\ |d_0| &> |d_{n-3}| \\ &\vdots \\ |r_0| &> |r_2| \end{aligned} \right\} (n-1) \text{ constraints} \quad (18)$$

It is noticed from Table 2 that a certain computation involved in the dotted entry in row  $2n-3$  becomes redundant. Therefore the corresponding second-order determinant to be calculated in entries  $2n-4$  and  $2n-5$  is also redundant.

\* For instance,  $A_k^2 - B_k^2 = A_{k-1} A_{k+1} - B_{k-1} B_{k+1}$ ,  $k = 2, 3, \dots, n-1$  and  $A_n^2 - B_n^2 = (A_{n-1} - B_{n-1})^2 F(1) F(-1)$ .

Table 2. A Simplified Stability Procedure Form

Row	$z^0$	$z^1$	$z^2$	$z^3$	$\dots z^{n-k}$	$\dots$	$z^{n-l}$	$z^n$
1	$a_0$	$a_1$	$a_2$	$a_3$	$\dots a_{n-k}$	$\dots$	$a_{n-1}$	$a_n$
2	$a_n$	$a_{n-1}$	$a_{n-2}$	$a_{n-3}$	$\dots a_k$	$\dots$	$a_1$	$a_0$
3	$b_0$	$b_1$	$b_2$	$\dots$	$\dots$	$\dots$	$b_{n-1}$	
4	$b_{n-1}$	$b_{n-2}$	$b_{n-3}$	$\dots$	$\dots$	$\dots$	$b_0$	
5	$c_0$	$c_1$	$c_2$				$c_{n-2}$	
6	$c_{n-2}$	$c_{n-3}$	$c_{n-4}$				$c_0$	
.	.	.	.				.	
.	.	.	.				.	
.	.	.	.				.	
$2n-5$	$s_0$	$s_1$	$s_2$	$s_3$				
$2n-4$	$s_3$	$s_2$	$s_1$	$s_0$				
$2n-3$	$r_0$	$r_1$	$r_2$					

The number of roots inside the unit circle, if none of the constraints in (18) are equal, is given by the negative number of the product

$$P_k = [|a_0| - |a_n|] [|b_0| - |b_{n-1}|], \dots, [|r_0^2 - r_2^2| - |r_0 r_1 - r_1 r_2|]$$

$$k=1, 2, \dots, n \quad (19)$$

with

$$P_1 = [|a_0| - |a_n|], P_2 = [|a_0| - |a_n|] [|b_0| - |b_{n-1}|], \dots \quad (20)$$

The difference between  $n$  and this number yields the number of roots outside the unit circle.

#### Example 2

$$F(z) = 0.0025 + 0.08z + 0.4126z^2 - 1.368z^3 + z^4, n=4 \quad (21)$$

Table 3. Stability Test for Example 2

Row	$z^0$	$z^1$	$z^2$	$z^3$	$z^4$
1	0.0025	0.08	0.4126	-1.368	1
2	1	-1.368	0.4126	0.08	0.0025
3	$\approx -1$	1.368	-0.4116	-0.0834	
4	-0.0834	-0.4116	1.368	-1	
5	0.9936	-1.402	0.5256		

For the stability test or determination of number of roots inside the unit circle; there is:

$$F(1) = 0.1271 > 0, F(-1) = 2.703 > 0 \quad (22)$$

$$0.0025 < 1, |a_0| < a_n \quad (23)$$

$$1 > 0.0834, |b_0| > |b_{n-1}| \quad (24)$$

$$0.9936 > 0.5256, |c_0| > |c_{n-2}| \quad (25)$$

The system is stable and therefore all the roots are inside the unit circle.

An alternative simplified form of the stability constraints given in eqn (18) can be obtained by noting that

$$\left. \begin{aligned} b_0 &= a_0^2 - a_n^2 < 0 \\ |b_0| &> |b_{n-1}| \end{aligned} \right\} \quad (26)$$

and

$$b_0 - b_{n-1} < 0, b_0 + b_{n-1} < 0 \quad (26a)$$

Similarly, for

$$\left. \begin{aligned} d_0 &= c_0^2 - c_{n-2}^2 > 0 \\ |d_0| &> |d_{n-3}| \end{aligned} \right\} \quad (27)$$

one can write

$$d_0 - d_{n-3} > 0, d_0 + d_{n-3} > 0 \quad (28a)$$

Similar relationships could also be obtained for the other constraints. The new stability criterion can be summarized as follows:

$n$ -even\*

$$\left. \begin{aligned} F(1) &> 0, \quad F(-1) > 0 \\ a_0 - a_n &< 0, \quad a_0 + a_n > 0 \\ c_0 - c_{n-2} &> 0, \quad c_0 + c_{n-2} > 0 \\ e_0 - e_{n-4} &> 0, \quad e_0 + e_{n-4} > 0 \\ \vdots & \\ s_0 - s_2 &> 0, \quad s_0 + s_2 > 0 \end{aligned} \right\} n\text{-constraints} \quad (28b)$$

$n$ -odd\*

$$\left. \begin{aligned} F(1) &> 0, \quad F(-1) < 0 \\ b_0 - b_{n-1} &< 0, \quad b_0 + b_{n-1} < 0 \\ d_0 - d_{n-3} &> 0, \quad d_0 + d_{n-3} > 0 \\ f_0 - f_{n-5} &> 0, \quad f_0 + f_{n-5} > 0 \\ \vdots & \\ r_0 - r_2 &> 0, \quad r_0 + r_2 > 0 \end{aligned} \right\} n-1\text{-constraints}$$

The above form can be further simplified by noting that each of the above constraints is divisible by the preceding constraints with alternate signs; i.e.  $c_0 - c_{n-2}$  is divisible by  $a_0 + a_n$ , etc. This property is utilized in the next section to show the equivalence between the table form and the determinant form previously obtained<sup>1</sup>. This simplification is particularly useful if the coefficients of  $F(z)$  are given otherwise than in numbers.

\* It should be noted that the stability constraints for  $n$ -even could also be written in an alternate form such that the starting constraints are  $b_0 - b_{n-1} < 0$ , and  $b_0 + b_{n-1} < 0$ . In this manner one can show that the final constraint is only  $s_0 - s_2 > 0$ . Similarly, for  $n$ -odd the final constraint is only  $r_0 - r_2 > 0$ . These forms could be used advantageously in the critical study of stability limit<sup>20</sup>.

### Relationship between the Table and Determinant Forms<sup>1, 23</sup>

In a preceding work a stability criterion in terms of the coefficients of  $F(z)$  is obtained in a determinant form. Now the relationship between the stability criterion introduced in this paper and the mentioned criterion will be established. The importance of this relationship lies in the fact that in order to obtain the former criterion<sup>1</sup> only second-order determinant evaluation is required.

To establish the identity between the  $A_k \pm B_k$  of the preceding criterion and the coefficients of the table, the following may be noted

$$a_0 \pm a_n = A_1 \pm B_1 \quad (29)$$

$$b_0 \pm b_{n-1} = A_2 \pm B_2 \quad (30)$$

The above is evident by direct examination of the terms of  $A_1 \pm B_1$  and  $A_2 \pm B_2$  and the corresponding terms of the table. Furthermore, by regular division the following relationships can be further identified

$$A_3 - B_3 = \frac{c_0 - c_{n-2}}{a_0 + a_n}, A_3 + B_3 = \frac{c_0 + c_{n-2}}{a_0 - a_n} \quad (31)$$

The above can also be written as

$$A_3 \pm B_3 = \frac{c_0 \pm c_{n-2}}{(A_1 \mp B_1)} \quad (32)$$

Similarly, by division and identifying the various terms, the following is obtained

$$A_4 \pm B_4 = \frac{d_0 \pm d_{n-3}}{(a_0^2 - a_n^2)(A_2 \mp B_2)} \quad (33)$$

Following the above procedure the following general formula can be obtained:

$$A_k \pm B_k = \frac{l_0 \pm l_{n-k+1}}{(a_0^2 - a_n^2)^{k-3} (b_0^2 - b_{n-1}^2)^{k-4}, \dots, (h_0^2 - h_{n-k+3}^2)(A_{k-2} \mp B_{k-2})} \quad (34)^*$$

By imposing the stability constraints on the coefficients of the table for  $n$ -odd and  $n$ -even, respectively, the constraints can be readily obtained on the  $A_k \pm B_k$  which can be seen to coincide with the preceding criterion<sup>\*\*</sup>. This will be illustrated for  $n = 5$  and  $n = 6$ , respectively.

$n = 5$ :

The stability constraints for this case in terms of the constraints in eqn (27b) are:

$$b_0 - b_4 < 0, b_0 + b_4 < 0 \quad (35)$$

$$d_0 - d_2 > 0, d_0 + d_2 > 0 \quad (36)$$

$$F(1) > 0, F(-1) < 0 \quad (37)$$

Noting that

$$A_2 \pm B_2 = b_0 \pm b_4, A_4 \pm B_4 = \frac{d_0 \pm d_2}{(a_0^2 - a_5^2)(A_2 \mp B_2)}$$

and  $b_0 = a_0^2 - a_5^2$ .

\* This relationship has also been verified without the use of division, but by using the induction method<sup>18</sup>. The  $l$ 's and  $h$ 's are the appropriate entries in Table 2.

\*\* By modifying Table 2, slightly,  $A_k \pm B_k$  can also be obtained directly without determinant expansion<sup>1</sup>.

The above constraints in terms of  $A$ 's and  $B$ 's can be written by noting from  $b_0 \mp b^4 < 0$  that  $b_0 < 0$ ; for  $n = 5$ , one obtains:

$$A_2 - B_2 < 0, A_2 + B_2 < 0 \quad (38)$$

$$A_4 - B_4 > 0, A_4 + B_4 > 0 \quad (39)$$

$$F(1) > 0, F(-1) < 0 \quad (40)$$

$n = 6$ :

In this case the constraints are:

$$a_0 - a_6 < 0, a_0 + a_6 > 0 \quad (41)$$

$$c_0 - c_4 > 0, c_0 + c_4 > 0 \quad (42)$$

$$e_0 - e_2 > 0, e_0 + e_2 > 0 \quad (43)$$

$$F(1) > 0, F(-1) > 0 \quad (44)$$

Using the relationship introduced in eqns (29), (31), and (34), one obtains

$$A_1 - B_1 < 0, A_1 + B_1 > 0 \quad (45)$$

$$A_3 - B_3 > 0, A_3 + B_3 < 0 \quad (46)$$

$$A_5 - B_5 < 0, A_5 + B_5 > 0 \quad (47)$$

$$F(1) > 0, F(-1) > 0 \quad (48)$$

Following the above procedure, the relationships for the general case can be readily formulated<sup>1</sup>.

### Summary of Stability Constraints for Low-order Systems

$n = 2$ :

$$a_0 < a_2$$

$$F(1) > 0, F(-1) > 0$$

$n = 3$ :

$$|a_0| < a_3$$

$$a_0^2 - a_3^2 < a_0 a_2 - a_1 a_3$$

$$F(1) > 0, F(-1) < 0$$

$n = 4$ :

$$\text{Form I } (1) a_0 - a_4 < 0, a_0 + a_4 > 0$$

$$(2) A_3 - B_3 > 0, A_3 + B_3 < 0$$

$$(3) F(1) > 0, F(-1) > 0$$

$$\text{Form II } A_2 - B_2 < 0, A_2 + B_2 < 0,$$

$$\text{where } A_2 = a_0^2 - a_4^2, B_2 = a_0 a_3 - a_1 a_4$$

$$A_3 - B_3 > 0,$$

$$\text{where } A_3 - B_3 = a_0^3 + 2 a_0 a_2 a_4 + a_1 a_3 a_4 - a_0 a_4^2$$

$$- a_0^2 a_4 - a_2 a_4^2 - a_0 a_3^2 - a_0^2 a_2 - a_1^2 a_4 + a_4^3 + a_0 a_1 a_3$$

$$F(1) > 0, F(-1) > 0$$

### Remarks on Use of Table

In this section certain properties of the tables are dealt with. In addition, certain simplifications of the use of the tables are illustrated.

### Simplification of the Tables

In certain cases the numerical values in the rows  $(2k + 1)$  could be very large or very small, so that multiplication or division by a certain constant  $\lambda$  is desirable. This can be easily accomplished without changing the results of the stability test. For instance, in Table 2 the elements of row  $(2k + 1)$  are divided (or multiplied) by  $\lambda$ ; then the elements of row  $(2k + 2)$  are also divided (or multiplied) by the same factor, and for stability the ratio of the first column of these rows is evidently independent of ' $\lambda$ '. Therefore, for the stability test or for obtaining information on the roots inside the unit, this modification will not alter the results.

### The Case $F(z)$ has Zeros on the Circle<sup>4, 22</sup> $|z| = 1$ , and no Reciprocal Zeros\*

In the above case a certain  $\delta_k$  in Table 1 becomes zero (or equivalently a certain  $b_0 = b_{n-1}$  in Table 2). To treat such a case, the following theorem is used: If for some  $k < n$ ,  $\delta_k \neq 0$ , but  $F_{k+1} \equiv 0$ , then  $F(z)$  has  $n - k$  zeros on the unit circle  $|z| = 1$  at the zeros of  $F_k(z)$ . It also has  $p$  zeros in the unit circle, where  $p$  is the number of negative  $P_j$ ,  $1 \leq j \leq k$ , and  $q = k - p$  zeros outside the circle.

### Example 3

Let  $F(z) = 1 + 2z + z^2 + 2z^3$

	$z^0$	$z^1$	$z^2$	$z^3$
$F(z)$	1	2	1	2
$F^*(z)$	2	1	2	1
$F_1(z)$	-3	0	-3	
$F_1^*(z)$	-3	0	-3	
$F_2(z)$	0	0		

It is noticed from the above that  $\delta_1 < 0$ ,  $F_2 \equiv 0$ , and no reciprocal zeros of  $F_1(z)$ ; thus one root is inside the unit circle and two roots are on the unit circle. The location of these two roots is obtained as follows:

$$F_1(z) = -3 - 3z^2 = -3(1 + z^2) = 0$$

therefore the roots are at  $z_{2,3} = \pm j$ .

\* The case when  $F(z)$  has both zeros on the unit circle and reciprocal zeros could be treated by a certain modification of this theorem<sup>22</sup>.

### The Case $F(z)$ has no Roots on the Unit Circle, but $\delta_{k+1} = 0$

In this case the number  $p$  of zeros of  $F(z)$  in the unit circle may be obtained by a limiting process as follows<sup>4</sup>:  $F_k(z)$  is replaced by the polynomial

$$f_k(z) = F_k(rz)$$

which, for  $r = 1 \pm \varepsilon$  and  $\varepsilon$  is a sufficiently small positive quantity, has as many zeros in the circle as does  $F_k(z)$ .

### Example

$$F(z) = -2 + 3z + 2z^2$$

In this case it is noticed that  $\delta_1 = a_0^2 - a_n^2 = 0$  (and  $F(z)$  has no zero on the unit circle); therefore,  $z$  in  $F(z)$  is replaced by  $rz = (1 - \varepsilon)z$ , with  $\varepsilon = 0.1$ , thus

$$f_k(z) = -2 + 2.7z + 1.62z^2$$

In the above case  $\delta_1 = 4 - (1.62)^2 > 0$ , and

$$f_k(1) > 0, f_k(-1) < 0, \text{ or } \delta_2 < 0$$

Therefore  $P_1 = \delta_1 > 0$ ,  $P_2 = \delta_1 \delta_2 < 0$ , one negative sign, hence one root exists outside the unit circle, and the other root inside the unit. It should be noted that the same results are obtained if  $\varepsilon$  is considered to be negative, i.e.  $r = 1 + \varepsilon = 1.1$ .

### Number of Roots Inside a Circle of Radius $\sigma < 1$

To obtain information on the number of roots inside a circle of radius  $\sigma < 1$ , one may apply a certain transformation on  $F(z)$  such as the circle of radius  $\sigma$ , becomes the unit circle. Hence, the proceeding method is readily applicable.

The above problem is useful in the relative stability test of the characteristic polynomial. The constraints for  $F(z) = \sum_{p=0}^n a_p z^p = 0$ , to have all its roots inside  $|z| = \sigma < 1$  will be obtained by replacing in the constraints found for  $\sigma = 1$  the following:

$$a_0 \rightarrow a_0, a_1 \rightarrow a_1 \sigma, \dots, a_p \rightarrow a_p \sigma^p, \dots, a_n \rightarrow a_n \sigma^n$$

For instance, those constraints for  $n = 2$  and  $n = 3$  are given as follows:

$$n = 2: \quad F_1(z) = a_0 + a_1 z + a_2 z^2, a_2 > 0$$

$$|a_0| < a_2 \sigma^2$$

$$F_1(\sigma) = a_0 + a_1 \sigma + a_2 \sigma^2 > 0$$

$$F_1(-\sigma) = a_0 - a_1 \sigma + a_2 \sigma^2 > 0$$

$$n = 3: \quad F(z) = a_0 + a_1 z + a_2 z^2 + a_3 z^3, a_3 > 0$$

$$|a_0| < a_3 \sigma^3$$

$$|a_0^2 - \sigma^6 a_3^2| > |(a_0 a_2 - \sigma^2 a_1 a_3)| \sigma^2$$

$$F_1(\sigma) = a_0 + a_1 \sigma + a_2 \sigma^2 + a_3 \sigma^3 > 0$$

$$F_1(-\sigma) = a_0 - a_1 \sigma + a_2 \sigma^2 - a_3 \sigma^3 < 0, \sigma \leq 1$$

As an extension of the above theorem, one can also obtain the stability conditions or the number of roots of a shifted unit circle. This is based on a theorem by Marden<sup>4</sup> given as follows.

To obtain the number of roots in the circle  $|z - s| < 1$ , replace the  $a_p$ 's of  $F(z)$  by the following:

$$a_p \rightarrow \sum_{k=p}^n C(k, p) s^{k-p} a_k$$

where the binomial  $C(k, p) = k!/p!(k-p)!$

In particular, the number of roots in the unit shifted circle<sup>6</sup> is obtained as shown in Figure 1, by letting  $s = 1$  in the above shifting transformation.

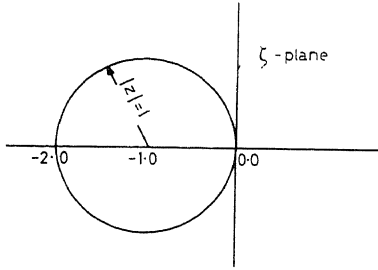


Figure 1. A shifted unit circle

#### Number of Real Roots of $F(z)$ between $-1$ and $+1$ in the $z$ plane

The above information is useful in locating the roots of the  $F(z)$  polynomial. This also has an effect on the synthesis of linear discrete systems when  $F(z)$  is stable. The procedure is based on 'Sturm's sequences'<sup>8</sup>. It is indicated in the following steps.

- (1) Consider  $F_1(z)$ , the derivative of  $F(z)$
- (2) Divide  $F_1(z)/F(z)$ , to obtain a remainder  $-F_2(z)$
- (3) Divide  $F_1/F_2$  to obtain a remainder  $-F_3$
- (4) Continue the division process until the remainder becomes a constant  $-F_n$
- (5) Consider the following two sequences:

$F(-1) F_1(-1) F_2(-1) F_{n-1}(-1) F_n$ , to obtain  $N_{-1}$  changes of sign, and  $F(1) F_1(1) F_2(1) F_{n-1}(1) F_n$ , to obtain  $N_{+1}$  changes of sign.

The number of real roots ' $\varepsilon$ ' of  $F(z)$  between  $(-1)$  and  $(+1)$  is given by

$$\varepsilon = N_{-1} - N_{+1}$$

The general conditions for the roots of  $F(z)$  to be real can be obtained using the early work of Hemite and the more recent work of Romanov<sup>13</sup>. In combination with the stability constraints one can obtain the necessary and sufficient condition for the roots of  $F(z)$  to be real and to lie between  $(-1)$  and  $(+1)$ , or between the origin and  $(+1)$  in the unit circle<sup>19</sup>.

#### Conclusions

It has been shown in this paper that one can easily obtain information on the number of roots of a real polynomial in the  $z$  plane. The method illustrated is based on the use of a table form, which consists of two different forms, Table 2 being the simpler form. A necessary and sufficient condition for all the roots to lie inside the unit circle (or equivalently, for the linear discrete system to be stable) may be obtained from the table.

The method which has been introduced herein is particularly useful if the coefficients of  $F(z)$  are given in numbers. It is also useful for design if one or two parameters are variable. An illustrative example is discussed in detail in Appendix II.

An alternate form of the table, which is based on a division method, is introduced in Appendix I. Both the table and the division method could easily be simulated on a digital computer for stability study of the characteristic polynomials.

The connection between the table and division method and the earlier introduced criteria of stability is established in this paper. The determinant method of the first criterion<sup>1</sup> is particularly useful for design, as is also shown in Appendix II. Furthermore, the evaluation of those higher-order determinants in the first criterion can be easily deduced from the table form which requires evaluation of only second-order determinants.

It is hoped that the earlier introduced criterion<sup>1</sup>, in combination with the discussion included herein, will finally solve the stability problem in a concise and satisfactory form. The future application by workers in this field will undoubtedly indicate the usefulness of these three stability criteria.

#### Appendix I — A Stability Test Using a Simple Division<sup>14</sup>

In this Appendix an alternate form of the stability test given in Table 2 is introduced. It is based on simple division of polynomials, and its derivation could be obtained from Table 2 or by using Rouché's Theorem directly. Since Table 2 is derived, it is used for obtaining the simple division method.

##### Stability Test

A necessary and sufficient condition for the real polynomial

$$F(z) = a_0 + a_1 z + a_2 z^2 + \dots + a_k z^k + \dots + a_n z^n, \text{ with } a_n > 0 \quad (49)$$

to have all its roots inside the unit circle in the  $z$  plane is obtained as follows:

$$F^*(z) = z^n F(1/z) = a_n + a_{n-1} z + a_{n-2} z^2 + \dots + a_{n-k} z^k + \dots + a_0 z^n \quad (50)$$

Obtain the stability constraints  $\alpha_k$  by simple division as shown:

$$\begin{aligned} \frac{F(z)}{F^*(z)} &\rightarrow \alpha_0 + \frac{F_1(z)}{F_1^*(z)} \rightarrow \alpha_0 + \alpha_1 + \frac{F_2(z)}{F_2^*(z)} \rightarrow \alpha_0 + \alpha_1 + \alpha_2 \\ &+ \frac{F_3(z)}{F_3^*(z)} \dots \rightarrow \alpha_0 + \alpha_1 + \alpha_2 + \dots + \alpha_{n-2} + \frac{F_{n-1}(z)}{F_{n-1}^*(z)} \end{aligned} \quad (51)$$

where  $F_1(z)$ ,  $F_2(z)$ , ... are the remainder polynomials obtained from the division and  $F_1^*(z)$ ,  $F_2^*(z)$ , ... are their inverses, as defined in (50). To obtain the inverse of  $F_1(z)$ , one factors out the common factor ' $z$ ' or its powers before applying (50). Similarly for  $F_2(z)$  and  $F_3(z)$ , ...

Satisfy the following constraints for the stability test:

$$\begin{aligned} F(1) &> 0, F(-1) > 0, n \text{ even} \\ &< 0, n \text{ odd} \\ |\alpha_k| &< 1, \quad k = 0, 1, 2, \dots, n-2 \end{aligned} \quad (52)$$

The proof of the test lies in the use of the preceding stability criterion, for it can be easily established that the identity of  $\alpha_k$ 's and the stability constants  $a_0, b_0, c_0, \dots$  of Table 2 are as follows.

$$\begin{aligned}\alpha_0 &= \frac{a_0}{a_n} \\ \alpha_1 &= \frac{b_{n-1}}{b_0} \\ \alpha_2 &= \frac{c_{n-2}}{c_0} \\ &\vdots \\ \alpha_{n-2} &= \frac{r_2}{r_0}\end{aligned}$$

From the stability constraints imposed previously on the  $a$ 's,  $b$ 's, ..., the constraints to be imposed on the  $\alpha_k$ 's are readily established as explained in (18).

The number of roots of  $F(z)$  inside the unit circle is equal to the number  $p$  which is equal to the negative products  $P_k$  defined as:

$$P_k = [|\alpha_0| - 1] \left[ \frac{1}{|\alpha_1|} - 1 \right], \dots, \left[ \frac{1}{|\alpha_k|} - 1 \right], \dots, \left[ \frac{1}{|\alpha_{n-1}|} - 1 \right]$$

$k = 1, 2, 3, \dots, n$

with

$$P_1 = [|\alpha_0| - 1], P_2 = [|\alpha_0| - 1] \left[ \frac{1}{|\alpha_1|} - 1 \right], \dots$$

and  $|\alpha_{n-1}|$  is obtained from,

$$\frac{F_{n-1}(z)}{F_{n-1}^*(z)} \rightarrow \alpha_{n-1} + \frac{F_n}{F_n^*}$$

#### Appendix II—Fourth Order Example<sup>1,15,17</sup>

In the sampled-data system shown in Figure 2, the use of the stability constraints discussed for the fourth-order case to obtain the maximum value of the gain ' $k$ ' for the stability is shown.

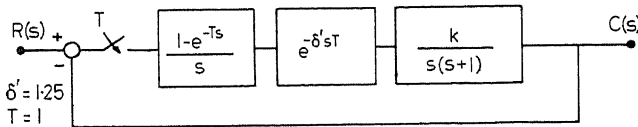


Figure 2. A sampled-data system with pure delay

#### The System Transfer Function

$$\frac{C^*(z)}{R^*(z)} = \frac{G_1^*(z)}{1 + G_1^*(z)}$$

where

$$G_1^*(z) = 0.2223 k \frac{(z+0.03)(z+1.755)}{z^2(z-1)(z-0.368)}$$

The characteristic polynomial  $1 + G_1^*(z) = 0$ , can be written as follows<sup>17</sup>:

$$z^4 - 1.368z^3 + (0.368 + 0.2223k)z^2 + 0.3974kz + 0.0123k = 0$$

In this case

$$\begin{aligned}a_4 &= 1, a_3 = -1.368, a_2 = 0.368 + 0.2223k, \\ a_1 &= 0.397k, a_0 = 0.0123k\end{aligned}$$

From Form II of the stability constraints for the fourth-order case,

$$A_2 + B_2 < 0, 0.000151k^2 - 0.4142k - 1 < 0, -2.4 < k < 2740$$

$$A_2 - B_2 < 0, 0.000151k^2 + 0.4142k - 1, -2740 < k < 2.4$$

$$A_3 - B_3 > 0, -(k-0.72)(k+5.5)(k+247) > 0, k < -247,$$

$$\text{or } -5.5 < k < 0.72$$

$$F(1) > 0, k > 0$$

$$F(-1) > 0, k < 16.75$$

From the above constraints on  $k$  it is evident that  $k_{\max} = 0.72$ .

From Form I,

$$|A_1| < B_1, |0.0123k| < 1, k < 81$$

$$A_3 - B_3 > 0, k < -247, \text{ or } -5.5 < k < 0.72$$

$$A_3 + B_3 < 0, -(k-6.27)(k-279.55)(k+1.37) < 0,$$

$$-1.37 < k < 6.27, \text{ or } k > 279.55$$

$$F(1) > 0, k > 0$$

$$F(-1) > 0, k < 16.75$$

From the above constraints again, as expected,  $k_{\max} = 0.72$ . It may be noted that the critical equation which gives the limiting gain in this case is  $A_3 - B_3 \geq 0$ . Thus in this case only the equations  $A_3 - B_3 = 0$  and  $F(1) = 0$ ,  $F(-1) = 0$  need be solved. Similarly, for higher-order systems which are initially stable and for which some parameters change so as to approach instability, only the equations  $A_{n-1} = B_{n-1}$ , and  $F(1) = 0$ ,  $F(-1) = 0$  are the most critical ones, which need be solved<sup>1, 20</sup>.

This research was supported by the United States Air Force Office of Scientific Research of the Air Research and Development Command, under Contract No. AF18 (600)-1521. The author is grateful to Mr. Jean Blanchard for his co-operation in the early stages of this work.

#### References

- JURY, E. I. A simplified stability criterion for linear discrete systems. Univ. of Calif., Berkeley, ERL Rep. Ser. 60, No. 373 (1961); *Proc. Inst. Radio Engrs N.Y.* 50, No. 6 (1962) 1493
- COHN, A. Über die Anzahl der Wurzeln einer Algebraischen Gleichung in einem Kreise. *Math. Z.* 14 (1922) 110
- FUJIIWARA, M. Über die Algebraischen Gleichungen, deren Wurzeln in einem Kreise oder in einer Halbebene liegen. *Math. Z.* 24 (1926) 160
- MARDEN, M. The geometry of the zeros of a polynomial in a complex variable. *Amer. Math. Soc.*, N.Y. (1949) 152
- SAMUELSON, P. H. Conditions that the roots of a polynomial be less than unity in absolute value. *Amer. Math. Statist.* 12 (1941) 360
- TSCHAUNER, J. Die Stabilität der Impulse-Systeme. *Regelungstechnik* 8 (1960) 42

- <sup>7</sup> JURY, E. I. Discussion on the stability of sampled-data systems. *Regelungstechnik* 8 (1961) 340
- <sup>8</sup> GANTMACHER, F. R. *The Theory of Matrices*. Vol. II. 1959. New York; Chelsea Publishing Co.
- <sup>9</sup> TSYPKIN, Y. *Theory of Pulse Systems*, pp. 423–427. 1958. Moscow (in Russian). State Press for Physics and Mathematical Literature
- <sup>10</sup> BROWN, B. M. *The Mathematical Theory of Linear Systems*, pp. 119–152. 1961. New York; Wiley
- <sup>11</sup> ROUCHE, E. Memoire sur la serie de Lagrange. *J. Ec. polyt. Paris* 22 (1862) 217
- <sup>12</sup> JURY, E. I. and BLANCHARD, J. A stability test for linear discrete systems in table form. *Proc. Inst. Radio Engrs, N.Y.* 44 (1961) 1947
- <sup>13</sup> ROMANOV, M. I. Algebraic criteria for aperiodicity of linear systems. *Sov. Phys. Dokl.* 4, No. 5 (1960) 955
- <sup>14</sup> JURY, E. I. A stability test for linear discrete systems using a simple division. *Proc. Inst. Radio Engrs* 44 (1961) 1948
- <sup>15</sup> JURY, E. I. Additions to notes on the stability criterion for linear discrete systems. *Trans. Inst. Radio Engrs, N. Y. (Automatic Control)*, Vol. AC-6 (1961) 242
- <sup>16</sup> JURY, E. I. and GUPTA, S. C. On the stability of linear discrete systems. *Regelungstechnik*, Pt. 4, No. 10 (1962) 157
- <sup>17</sup> JURY, E. I. *Sampled-data Control Systems*. 1958. New York; Wiley
- <sup>18</sup> JURY, E. I. Proof of a general relationship used in the stability of test of linear discrete systems, Addendum. Univ. of Calif., Berkeley, ERL Rep. Ser. 60, No. 373 (June 1962)
- <sup>19</sup> JURY, E. I. and PAVLIDIS, T. Aperiodicity Criteria for linear discrete systems, *Inst. Radio Engrs, PGCT*, Vol. CT-9, No. 4 (Dec. 1962) 431–433
- <sup>20</sup> JURY, E. I. and PAVLIDIS, T. Stability and aperiodicity constraints for system design. *Inst. Radio Engrs, N.Y. PGCT* (in the press)
- <sup>21</sup> WILF, H. S. A stability criterion for numerical integration. *J. Assoc. Comp. Mach.* 6 (1959) 363
- <sup>22</sup> JURY, E. I. *The Theory and Application of the Z-transform*, Classnotes 290-E Univ. of Calif. (September 1961) to be published 1964. New York; Wiley
- <sup>23</sup> JURY, E. I. On the generation of the stability constraints for linear discrete systems. To be published by the *I.E.E.E. P.G.A.C.* (April 1963)

## DISCUSSION

M. HAMZA, *ETH Zurich, Switzerland*

Determining the stability of feedback control systems is a problem of major importance, consequently it is very useful to have efficient methods for checking system stability. This work is a definite contribution in that direction; it is lucidly written and the stability criteria presented are simple and practicable.

Professor Jury considers real polynomials and mentions that some of the theorems he gives can be applied to the general coefficient case. Could Professor Jury please indicate which theorems he is referring to? Further, if the coefficients are of the general type which methods for investigating system stability does he recommend? Does the method recommended cover systems having a pure delay?

E. I. JURY, *in reply*

It may be indicated that the theorem quoted after eqn (4) is applicable to the general coefficient case. However, in this case eqn (3) should be modified to read

$$a_k^{(j+1)} = a_0^{(j)} \bar{a}_k^{(j)} - \bar{a}_{n-j}^{(j)} a_{n-j-k}^{(j)}$$

where  $\bar{a}_k$  is the complex conjugate of  $a_k$ .

Table 1 is also applicable to stability tests for the general coefficients case, provided the entries of the second, fourth, ... rows are the complex conjugate of the first, third, ... rows. If the pure delay in the system is approximated by rational function of  $z$ , then the stability criterion introduced is equally applicable.

B. M. BROWN, *Royal Naval College, Greenwich, London S.E. 10.*

The development of linear discrete systems and digital processes in the last 20 years, has led to an increase in importance of the corresponding stability theory. This theory was initiated some 40 years ago by Schur, Cohn and others, but until recently the only easily available account of it seems to have been in Marden's monograph. This is concerned primarily with polynomials with complex coefficients and is therefore more complicated than is necessary for a discussion of the practical problems of stability. Marden made certain original contributions, but since then not much work seems to have been done on the problem which, in view of its fundamental nature, is rather surprising. Professor Jury is therefore to be commended for undertaking his detailed and exhaustive study. Some months ago he was good enough to send me some of his work, since when this fascinating problem has been my

major preoccupation. My main results have been a new method of deriving with the singular cases. In addition, I have used similar techniques to prove the Routh and Hurwitz criteria and their various extensions. These proofs are shorter and simpler than any others I have come across and I hope to publish them shortly.

When the characteristic equation of a continuous or discrete linear system is given the discussion of its stability can take two forms. First, if numerical values are given for all the coefficients the simplest procedure with a continuous system is, in general, to use Routh's method. The corresponding process with a discrete system is that described in Table 1 of the paper, or the equivalent and slightly simpler process given in Appendix I. If, however, it is desired to consider the effect of varying one or more of the coefficients or to carry out a general theoretical investigation, it is usually better with a continuous system to use the determinant sequence of Hurwitz. The equivalent determinants for a discrete system are those which Professor Jury denotes by  $A_k \pm B_k$ , but which, unfortunately, he has been unable to quote in this paper.

E. I. JURY, *in reply*

Because of space limitation, I was unable to discuss in detail the formulation of  $A_k \pm B_k$ . However, these are discussed in detail in References 1 and 23 of the paper. One of the major objectives of this paper is to establish the connection between the table form, the division method and the determinant method. I look forward to reading Professor Brown's contribution on this general problem.

P. M. PARKS, *Department of Aeronautics, The University, Southampton*

Given the polynomial

$$P(z) \equiv z^n + a_1 z^{n-1} + a_2 z^{n-2} + \dots + a_n$$

the Schur-Cohn criterion states that all the zeros of  $P(z) = 0$  lie within the unit circle  $|z| = 1$  if, and only if, the principal minors of the matrix  $B$

$$B_{ij} = \sum_{p=0}^{\min(i,j)} (a_{i-p} a_{j-p} - a_{n+p-i} a_{n+p-j}) \quad (i, j = 0, 1, 2, \dots, n-1)$$

where  $a_0 \equiv 1$

are all positive, that is  $\mathbf{x}' B \mathbf{x}$  is a positive-definite quadratic form.





By the application of Mitrović's method<sup>1,2</sup>, it is convenient to map the damping region of Figure A into the plane of any of variable coefficients, say  $a_p$  and  $a_q$  ( $n \geq p > q \geq 0$ ), of the characteristic equation

$$F(z) = a_n z^n + a_{n-1} z^{n-1} + \dots + a_p z^p + \dots + a_q z^q + \dots + a_1 z + a_0 = 0 \quad (1)$$

By considering the coefficients  $a_p$  and  $a_q$  as variables  $\xi$  and  $\eta$ , respectively, and using the relationship

$$z = e^{sT} = \exp(-\omega_n \zeta + j\omega_n \sqrt{1-\zeta^2}) T \quad (2)$$

where  $\zeta$  is the relative damping coefficient,  $\omega_n$  is the undamped natural frequency and  $T$  is the sampling period, one obtains the following equations<sup>2</sup>

$$\begin{aligned} \xi &= \frac{\sum_{k=0}^n a_k e^{-(k-p)\omega_n \zeta T} \phi_{k-p}(-\cos \omega_n T \sqrt{1-\zeta^2})}{\sum_{k \neq p} a_k e^{-(k-p)\omega_n \zeta T} \phi_{k-p}(-\cos \omega_n T \sqrt{1-\zeta^2})} \\ \eta &= \frac{\sum_{k=0}^n a_k e^{-(k-q)\omega_n \zeta T} \phi_{k-q}(-\cos \omega_n T \sqrt{1-\zeta^2})}{\sum_{k \neq q} a_k e^{-(k-q)\omega_n \zeta T} \phi_{k-q}(-\cos \omega_n T \sqrt{1-\zeta^2})} \end{aligned} \quad (3)$$

In the  $0\xi\eta$  plane, the eqns (3) represent the loci of points corresponding to the roots with constant damping coefficient  $\zeta$ . If the region shown in Figure A is mapped into the  $0\xi\eta$  plane, using the eqns (1), (2) and (3), it is possible to determine all values of the variables  $\xi$  and  $\eta$  which correspond to the degree of relative stability indicated by  $M$  in Figure A.

#### References

- MITROVIĆ, D. Graphical analysis and synthesis of feedback control systems, III Sampled-data feedback control systems. *Trans. A.I.E.E.* 77, Pt. II (1958) (Jan. 1959 Sect.) 497-503
- ŠILJAK, D. Generalization of Mitrović's method. *I.E.E.E. Summer Gen. Meet.*, Toronto No. 63-988 (1963)

E. I. JURY, in reply

I might indicate that on page 147, under the heading 'Number of Roots Inside a Circle of Radius  $\sigma < 1$ ', one can use these results to obtain analytic conditions on relative stability of a different form than the one proposed. From the form indicated in Figure A of Dr. Šiljak's discussion, one can obtain analytic conditions, but these become quite complicated. In this case the graphical procedure indicated by Dr. Šiljak becomes more convenient, although quite involved graphically.

M. THOMA, *Institut für Regelungstechnik, Technische Hochschule Darmstadt, Darmstadt, Heidelberger Str. 11, West Germany*

In his interesting paper, Professor Jury deals mainly with the problem of testing the stability of linear sampled-data systems. The advantage of the given reduction method is that the bi-linear transformation, as the solution of determinants of higher than second order, is also avoided. But the theorem of Marden, on the right-hand side of page 143, and therefore eqn (7), is only valid if none of the  $\delta_k$  ( $k = 1, 2, \dots, n$ ) is missing. However, this case may easily occur with stable systems because it is not necessary that all the coefficients  $a_k$  ( $k = 1, 2, \dots, n$ ) of the equation  $F(z) = 0$  exist. Therefore, the statement (a) on page 143 is not quite general. It must be read: If  $F_j(z)$  is of degree  $n-j$  then  $F_{j+1}$  is at highest degree of  $n-j-1$ . For example, if

$$F(z) = a_0 + a_2 z^2 + a_5 z^5$$

then

$$F_1(z) = a_0^2 - a_5^2 + a_2 a_0 z^2 - a_2 a_5 z^3$$

which is of degree three and  $\delta_2$  does not exist.

In my paper<sup>1</sup> I gave almost simultaneously, but independent of Dr. Jury, another derivation of this result. It is more direct and there can be found a somewhat more uniform consideration of the problem. If we define eqn (1), not as it is done in the paper by Dr. Jury, but in the following manner:

$$F_{j+1}(z) = \frac{1}{z} [a_{n-j}^{(j)} F_j(z) - a_0^{(j)} F_j^*(z)], \quad j=0, 1, \dots, n-1$$

then the constant term of the polynomial in the squared brackets is zero. However, as proved by Cohn<sup>2</sup> and Thoma<sup>1</sup>, the polynomials  $F_j(z) = 0$  and  $z \cdot F_{j+1}(z) = 0$  do have for

$$|a_{n-j}^{(j)}| > |a_0^{(j)}|$$

the same number of roots inside the unit circle. Because the equation  $z \cdot F_{j+1}(z) = 0$  has the roots  $z = 0$  in the unit circle we have, for stable systems, just to show that the polynomial  $F_{j+1}(z)$  of degree  $n-j-1$  has only roots with  $|z| < 1$ . Thereby eqn (3) changes to

$$a_{n-1-j-k}^{(j+1)} = a_{n-j-k}^{(j)} \cdot a_{n-j-k}^{(j)} - a_0^{(j)} \cdot a_k^{(j)} \quad (k=0, 1, \dots, n-1-j)$$

$j+1$  is the number of reductions (e.g. for  $j=0$  we get the  $n-1$  coefficients of the first reduction, etc.). But all the roots of the polynomial can then only lie in the unit circle if the absolute value of the coefficient of the term with the highest power  $a_n$  (highest coefficient) is larger than that of the constant term  $a_0$  (last coefficient), i.e.

$$|a_n| > |a_0|$$

It is easily shown that for

$$|a_n| \leq |a_0|$$

not all the roots can lie inside the unit circle. We therefore divide the equation

$$F(z) = a_0 + a_1 z + \dots + a_{n-1} z^{n-1} + a_n z^n = 0$$

by  $a_n$ ; it yields

$$\left| \frac{a_0}{a_n} \right| \geq 1$$

because we assumed that

$$|a_n| \leq |a_0|$$

We split this equation into the factors

$$F(z) = (z - \gamma_n)(z - \gamma_{n-1}) \dots (z - \gamma_1) = 0$$

where  $\gamma_i$  are the  $n$  roots of  $F(z)$ . After multiplying out the last equation, the coefficient of  $z^n$  is equal to 1 and that of the constant term to

$$(-1)^n \gamma_n \cdot \gamma_{n-1} \dots \gamma_1$$

For

$$|a_n| \leq |a_0|$$

it is necessary that

$$\left| \frac{a_0}{a_n} \right| = |\gamma_n \cdot \gamma_{n-1} \dots \gamma_1| \geq 1$$

However, this is just possible if either all  $|\gamma_i| = 1$  or at least one root has an absolute value greater than 1. For a stable system it is therefore necessary that in the given equation  $F(z) = 0$ , as in all reduced equations,  $F_j(z) = 0$  the absolute value of the highest coefficients is in contrast to eqn (18) larger than that of the last coefficient

$$|a_{n-j}^{(j)}| > |a_0^{(j)}|, \quad j=0, 1, \dots, n-1$$

Besides this, the description<sup>1</sup> given can easily be extended to determine the number of roots inside, outside, and on the unit circle, and to polynomials with complex coefficients which is important for the determination of the root distribution<sup>3</sup>. As we discovered, the given procedures<sup>1,3</sup> are also very suitable for the automatic calculation of higher order polynomials with a digital computer. Further study of this problem can be found in the discussions<sup>4,5</sup>.

#### References

- <sup>1</sup> THOMA, M. Ein einfaches Verfahren zur Stabilitätsprüfung von linearen Abtastsystemen. *Regelungstechnik*, 10 (1962) 302-306
- <sup>2</sup> COHN, A. Über die Anzahl der Wurzeln einer algebraischen Gleichung in einem Kreise. *Math. Z.*, 14 (1922) 110-148
- <sup>3</sup> THOMA, M. Über die Wurzelverteilung von linearen Abtastsystemen. *Regelungstechnik*, 11 (1963) 70-74
- <sup>4</sup> JURY, E. I. Remarks on the paper 'Ein einfaches Verfahren zur Stabilitätsprüfung von linearen Abtastsystemen' by M. Thoma. *Regelungstechnik*, 10, No. 12 (1962) 554-555; Author's reply, *ibid.*, 11, No. 1 (1963) 30
- <sup>5</sup> JURY, E. I. Remarks on the paper 'Über die Wurzelverteilung von linearen Abtastsystemen' by M. Thoma. *Regelungstechnik*, 11, No. 7 (1963) 316-317; Author's reply, *ibid.*, 11, No. 8 (1963) 365-366

#### E. I. JURY, in reply

(1) The theorem of Marden, quoted after eqn (4), is quite correct. In some cases when certain coefficients of the polynomial  $F(z)$  do not appear, then some of the  $\delta_k$  do not exist. For such a special case the modification of Dr. Thoma is quite correct. This modification has already been incorporated in the page-proofs of the paper. These special cases are discussed in detail in References 2 and 22.

(2) In my earlier contribution<sup>12,14</sup> which slightly preceded Dr. Thoma's contribution, I have briefly presented the table and division method. In this paper the proofs and extension are indicated. In my Table 2 the stability test is simpler than Dr. Thoma's method because of the use of the auxiliary constraints  $F(1) > 0$  and  $(-1)^n F(-1) > 0$  as indicated in eqn (17). This fact and other points are discussed with Dr. Thoma in References 4 and 5 of his discussion. Furthermore, if the first constraint in eqn (18) is written as  $a_n > |a_0|$ , then our first  $(n-1)$  constraints become identical.

(3) In the present contribution, I have also discussed the root distribution within the unit circle as well as the stability condition. The singular cases are briefly discussed, and in Reference 22 a detailed study will be available. Finally, in my discussions with Dr. Thoma's contribution much light has been shed on the solution of this interesting problem, and the reader's attention has been drawn to the contribution of European and American authors in the past.

YA. Z. TSYPKIN, *Institute of Automation and Telemechanics, Moscow I-53, Kalanchevskaya 15-A, U.S.S.R.*

Professor Jury's paper gives an exhaustive solution to problems of determining the stability and the degree of stability of sampled-data systems, and it is difficult to add anything to the results obtained. I should merely like to make a few comments on this field of questions.

Algebraic criteria of stability are obviously suitable for characteristic equations of low powers, since the coefficients of these equations are linked by complex relationships to the parameters of the sampled-data system.

For high-order sampled-data systems we can speak of determining the stability of the system according to numerically-specified parameters.

The simplest way to determine stability seems to be by the tabular form<sup>1</sup>, corresponding to Schur's criterion, or the modifications of it given in Jury's paper. It would be desirable to compare all the numerical methods of stability determination from the point of view of the amount of computing operations.

The use of the Schur theorem for determining the number of real roots, lying on the segment  $-1, +1$ , involves cumbersome computations. They would be justified if such distribution of the roots had a particular purpose. I should like to ask the author when the need arises for such distribution of the roots.

#### Reference

- <sup>1</sup> TSYPKIN, YA. Z. Theory of Linear Sampled-data Systems, Moscow, 1963

#### E. I. JURY, in reply

The discussion on the number of real roots of  $F(z)$  between  $-1$  and  $+1$  in the  $z$  plane is mainly of academic interest. However, the conditions on the roots to lie between  $(0, 1)$  in the unit circle is of practical significance. It is related to the problem of aperiodicity or, alternately, for the sampled-data system response type of finite oscillations. This type of system is used in instrumentation feedback servos where excessive oscillations are not desirable. A thorough discussion of this aperiodicity problem is indicated in Reference 19 of the paper. As for comparison of the various methods for numerical test of stability, my investigation has shown that both the table form and the determinant method (provided the latter is written in the proper form) are equivalent<sup>1</sup>. However, for hand computation, the table method is more convenient.

#### Reference

- <sup>1</sup> JURY, E. I. On the evaluation of the stability determinants in linear discrete systems. *Inst. Radio Engrs. N.Y. P.G.A.C.* (July 1962)

# Analytical Approaches to Non-linear Sampled-data Control Systems

B. KONDO and S. IWAI

## Summary

The analysis of non-linear oscillations in sampled-data control systems is treated. The system considered here has the non-linear element in discrete signal portions. In the first half of the paper the strict analytical method is proposed. By using it, the periodic equilibrium states and their stability can be strictly analysed, without any assumption on the waveform, if the period of the oscillation is an integer multiple of the sampling period.

In the second half, two different applications of describing function methods to the sampled-data systems are proposed. One is defined in the same way as in continuous systems, and it is accurately applicable to the sinusoidal oscillation of which the period is sufficiently larger than the sampling period. The other is defined by considering the phase relation of oscillations at sampling instants, and it is applicable when both frequencies are closely approached.

Moreover, in some examples illustrated, the existence of 'hard self-excitations' in sampled-data systems is suggested.

## Sommaire

L'auteur expose l'analyse des oscillations non-linéaires dans des systèmes de commande du type à données échantillonnées. Dans le système considéré ici, l'élément non-linéaire réside dans le signal qui est échantillonné par portions discrètes. La première partie de la communication traite de la méthode strictement analytique. Cette méthode permet d'analyser strictement, sans aucune hypothèse sur la forme des oscillations, les états d'équilibre périodiques et leur stabilité, cela pour autant que la période des oscillations soit un multiple entier de la période d'échantillonnage.

Dans la seconde partie, on suggère deux applications différentes des méthodes utilisant une fonction descriptive à des systèmes à données échantillonnées. Pour l'une, la méthode est utilisée de la même manière que dans un système continu: elle s'applique de façon rigoureuse si l'oscillation est sinusoïdale et a une période suffisamment grande par rapport à la période d'échantillonnage. Pour l'autre, il faut tenir compte du déphasage entre les oscillations et l'échantillonnage: elle s'applique lorsque ces deux fréquences sont voisines.

En outre, l'existence possible d'auto-excitations dures (hard self-excitations) dans les systèmes à données échantillonnées sera signalée à propos de quelques-uns des exemples traités.

## Zusammenfassung

Der Aufsatz behandelt die Untersuchung nichtlinearer Schwingungen in Abtastregelsystemen. In dem hier betrachteten System tritt die Nichtlinearität während der diskreten Signalschnitte auf. Der erste Teil des Beitrages behandelt die strenge analytische Methode. Mit ihrer Hilfe lassen sich die periodischen Gleichgewichtszustände und ihre Stabilität genau untersuchen, wobei Annahmen über die Wellenform nicht notwendig sind, wenn nur die Schwingungsperiode ein ganzes Vielfaches der Abtastperiode ist.

Im zweiten Teil werden zwei verschiedene Anwendungen der Beschreibungsfunktionen auf Abtastsysteme vorgeschlagen. Eine ist in gleicher Weise wie bei den kontinuierlich arbeitenden Systemen definiert und kann mit guter Genauigkeit auf sinusförmige Schwingungen mit einer gegenüber der Abtastperiode genügend großen Periodendauer angewendet werden. Die andere betrachtet die Phasenlage der

Schwingungen in den Abtastaugenblicken; sie ist dann anwendbar, wenn beide Frequenzen nahe beieinander liegen.

Einige Beispiele lassen zudem vermuten, daß auch in Abtastsystemen „harte“ selbsterregte Schwingungen auftreten.

## Introduction

The analysis of a sampled-data control system with non-linearities is fairly difficult due to the presence of discontinuities of the signal in addition to the non-linear elements, such as saturation, backlash, relay, etc. However, the same methods as for continuous systems may be applicable to sampled-data systems which contain a non-linear element in continuous signal portions, if they are slightly modified, for example, by introducing the  $z$  transformation. Thus, a peculiarity of the analysis for non-linear sampled-data systems occurs when non-linearity exists in the discrete signal portion, because the signals in this portion contain many higher harmonic frequency components. This difficulty is more pronounced when the frequency of the non-linear oscillation approaches the sampling frequency.

From this viewpoint, such systems as indicated in *Figure 1 (a)* and *(b)* are considered in this paper. If the hold circuit  $H(s)$  is a zero-order hold, the phenomena observed in the circuits *(a)* and *(b)* are evidently identical.

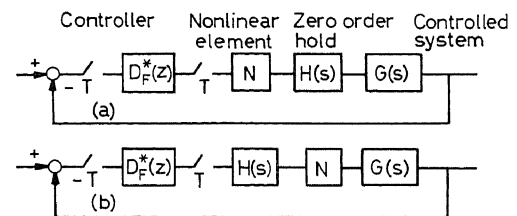


Figure 1. Non-linear sampled-data control system

## Analysis of Non-linear Oscillation in a Sampled-data Control System by the Method Without any Approximation

If both the input and the output of a non-linear element are expressed in the form of pulse series  $x^*(t)$  and  $y^*(t)$ , respectively, in *Figure 2* the following relations hold:

$$x^*(t) = \sum_{i=0}^{\infty} x_i \delta(t - iT) \quad (1)$$

$$y^*(t) = \sum_{i=0}^{\infty} y_i \delta(t - iT) = \sum_{i=0}^{\infty} \alpha_i x_i \delta(t - iT)$$

where  $\alpha_i$  is a weighting coefficient for the input pulses and varies with the characteristics of the non-linear element and the

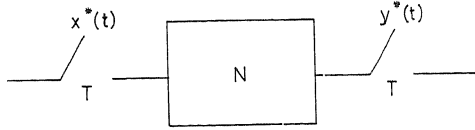


Figure 2. Input and output of non-linear element

magnitude of each value of  $x_i$ . When the input  $x^*(t)$  is a periodic pulse series  $\tilde{x}^*_t$  of period  $lT$ , where  $l$  is an integer, then the  $l$  kinds of  $\alpha_i$  ( $i = 1 \sim l$ ) are defined corresponding to  $l$  kinds of input pulses  $x_1, x_2, \dots, x_l$ , respectively. By applying this consideration to the non-linear sampled-data system indicated in Figure 3 (a), it can be easily found that the equilibrium state for the periodic oscillation of period  $lT$ , can be treated on the equivalent circuit shown in Figure 3 (b), where the rotary switch  $S$  is successively connected with the elements  $\alpha_i$  in the order of  $\alpha_1, \alpha_2, \dots, \alpha_l, \alpha_1, \dots$  with the time interval  $T$ .  $x_i$  denotes the value of the error signal at the instant when  $\alpha_i$  is connected.

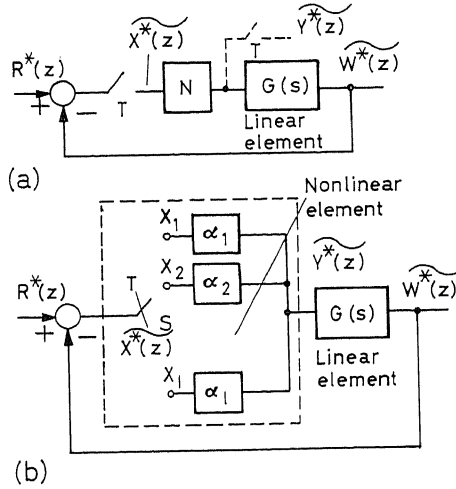


Figure 3. Sampled-data system: (a) original system; (b) equivalent system

### Equilibrium States

The  $z$  transforms of  $\tilde{x}^*(t)$  and  $\tilde{y}^*(t)$  are represented by

$$\begin{aligned} \tilde{x}^*(z) &= \sum_{i=1}^l \frac{x_i z^{-(i-1)}}{(1-z^{-l})} \\ \tilde{y}^*(z) &= \sum_{i=1}^l \frac{y_i z^{-(i-1)}}{(1-z^{-l})} = \sum_{i=1}^l \frac{\alpha_i x_i z^{-(i-1)}}{(1-z^{-l})} \end{aligned} \quad (2)$$

Assuming the system to be in self-oscillation and the steady-state value of the system input  $r(t)$  to be zero, gives the relation

$$\tilde{X}^*(z) = -\tilde{W}^*(z) = -G^*(z) Y^*(z) \quad (3)$$

where  $W^*(z)$  is the  $z$  transform of the system output  $w(t)$  and  $G^*(z)$  is the pulse transfer function of the linear element  $H(s)G(s)$ . In general,  $G^*(z)$  is represented by

$$G^*(z) = \frac{a_0 + a_1 z^{-1} + \dots + a_m z^{-m}}{1 + b_1 z^{-1} + \dots + b_m z^{-m}} \quad (4)$$

From eqns (2)–(4), for the equilibrium state of period  $lT$ , the following relations are obtained:

$$\begin{aligned} (B_1 + A_1 \alpha_l) x_l + (B_2 + A_2 \alpha_{l-1}) x_{l-1} + \dots + (B_l + A_l \alpha_1) x_1 &= 0 \\ (B_2 + A_2 \alpha_l) x_l + (B_3 + A_3 \alpha_{l-1}) x_{l-1} + \dots + (B_1 + A_1 \alpha_1) x_1 &= 0 \\ \vdots & \\ (B_l + A_l \alpha_l) x_l + (B_1 + A_1 \alpha_{l-1}) x_{l-1} + \dots + (B_{l-1} + A_{l-1} \alpha_1) x_1 &= 0 \end{aligned} \quad (5)$$

where  $B_i$  and  $A_i$  ( $i = 1 \sim l$ ) are

$$\begin{aligned} B_1 &= 1 + b_1 + b_{2l} + \dots & A_1 &= a_0 + a_l + a_{2l} + \dots \\ B_2 &= b_1 + b_{l+1} + b_{2l+1} + \dots & A_2 &= a_1 + a_{l+1} + a_{2l+1} + \dots \\ \vdots & & \vdots & \\ B_l &= b_{l-1} + b_{2l-1} + b_{3l-1} + \dots & A_l &= a_{l-1} + a_{l-1} + a_{2l-1} + \dots \end{aligned} \quad (6)$$

If the non-linearity is, for example, the saturation shown in Figure 4, then

$$\alpha_i = 1 \text{ for } |x_i| \leq \Delta, \quad \alpha_i = \frac{\Delta}{x_i} \text{ for } x_i \geq \Delta, \quad \alpha_i = -\frac{\Delta}{x_i} \text{ for } x_i \leq -\Delta \quad (7)$$

As shown in this example, when the non-linearity is a single-valued function of  $x_i$ , then the weighting coefficients are uniquely defined as the functions of  $x_i$ . Therefore, eqns (5) have  $l$  unknown quantities,  $x_i$  or  $\alpha_i$  ( $i = 1 \sim l$ ), and the number of unknowns coincides with the number of equations. In other words, eqns (5) give the equilibrium conditions for the non-linear oscillation of period  $lT$  in the system with the non-linearity defined by a single-valued function of the input pulses.

Now, if each value of input pulses  $x_i$  is assumed, for example, as  $x_1 \geq \Delta$ ,  $|x_2| < \Delta$ ,  $x_3 \leq -\Delta$ ,  $\dots$ , then substituting eqn (7) into eqn (5), yields a set of the first order simultaneous equations of  $x_i$ . When the solutions of the simultaneous equations satisfy eqn (7), the equilibrium state is established. On the other hand, if they do not satisfy, the equilibrium state as assumed above does not exist. These procedures must be applied for all sets of possible combinations of  $x_i$ . For example, when  $l = 2, 3$  and  $4$ , the number of such sets is 2, 4 and 10, respectively. Thus, in general, the number of sets increases as  $l$  becomes larger, but the difficulty of the analysis is fortunately independent of the value of  $m$ , the order of the system. However,

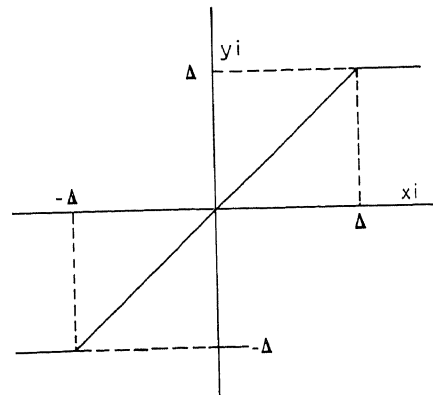


Figure 4. Saturation element

any complementary condition such that, for example, the system has a pole at the origin of a complex plane, will be very effective in decreasing the number of sets.

#### Stability of the Equilibrium State

In general, the sampled output of a non-linear element can be expressed in the form:

$$\tilde{y}^*(t) \equiv y^*[\tilde{x}^*(t)] \quad (8)$$

Introducing this notation into eqn (3) gives

$$Z[\tilde{x}^*(t)] = -G^*(z) Z\{\tilde{y}^*[\tilde{x}^*(t)]\} \quad (9)$$

Assuming that the periodic equilibrium state is affected by small non-periodic disturbances, the input of the non-linear element  $\tilde{x}^*(t)$  is superposed by

$$\delta x^*(t) = \delta x_0 \delta(t) + \delta x_1 \delta(t-T) + \delta x_2 \delta(t-2T) + \dots \quad (10)$$

Then the sampled output of the non-linear element can be expressed by  $y^*[\tilde{x}^*(t) + \delta x^*(t)]$  and eqn (9) becomes

$$Z[\tilde{x}^*(t) + \delta x^*(t)] = -G^*(z) Z\{y^*[\tilde{x}^*(t) + \delta x^*(t)]\} \quad (11)$$

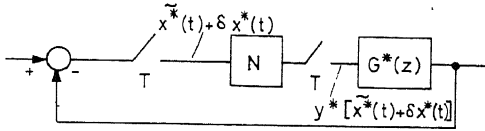


Figure 5. System used for stability criterion

Such a system is shown in Figure 5. Since clearly

$$Z[\tilde{x}^*(t) + \delta x^*(t)] = Z[\tilde{x}^*(t)] + Z[\delta x^*(t)] \quad (12)$$

substituting eqns (9) and (12) into (11) yields

$$Z[\delta x^*(t)] = -G^*(z) Z\{y^*[\tilde{x}^*(t) + \delta x^*(t)] - y^*[\tilde{x}^*(t)]\} \quad (13)$$

or

$$Z[\delta x^*(t)] = -G^*(z) Z\left\{\frac{y^*[\tilde{x}^*(t) + \delta x^*(t)] - y^*[\tilde{x}^*(t)]}{\delta x^*(t)} \delta x^*(t)\right\} \quad (14)$$

The extreme value of this equation as the disturbance becomes infinitesimal gives

$$Z[\delta x^*(t)] = -G^*(z) Z\{y^*[\tilde{x}^*(t)] \delta x^*(t)\} \quad (15)$$

where  $y^*[\tilde{x}^*(t)]$  is the derivative of  $y^*[\tilde{x}^*(t)]$  with the input pulse series  $\tilde{x}^*(t)$ , where  $\tilde{x}^*(t)$  is a periodic signal with period  $lT$  and can be represented by the repetition of  $l$  pulses  $x_1 \sim x_l$ . When the non-linearity is, for example, the saturation shown in Figure 4,  $y^*$  is represented as

$$y^*[\tilde{x}^*(t)]|_{x_i} = \xi_i, \quad \xi_i = 0 \text{ for } |x_i| > \Delta \\ \xi_i = 1 \text{ for } |x_i| < \Delta \quad (16)$$

Therefore,

$$\dot{y}^*[\tilde{x}^*(t)] \delta x^*(t) = \dots \\ + \xi_1 \delta x_1 \delta(t-T) + \xi_2 \delta x_2 \delta(t-2T) + \dots + \xi_l \delta x_l \delta(t-lT) \\ + \xi_1 \delta x_{l+1} \delta(t-l+1T) + \xi_2 \delta x_{l+2} \delta(t-l+2T) + \dots \quad (17)$$

Thus the stability problem of the equilibrium state can be converted to that of the sampled-data system with the periodically varying gain illustrated by Figure 6.

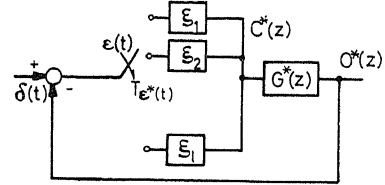


Figure 6. Variable gain sampled-data system

Since the system shown in Figure 6 is a kind of linear system, the stability can be determined by the convergency of the response to a unit impulse disturbance. Denoting the value of the impulse response of the linear system at  $t = nT$  by  $g_n$ ,

$$g_n = \frac{1}{2\pi j} \int G^*(z) z^{n-1} dz \quad (18)$$

Now, if

$$G^*(z) \equiv \frac{N(z)}{M(z)}, \quad M(z) = (z-r_1)(z-r_2)\dots(z-r_m) \quad (19)$$

$r_1 \neq r_2 \neq \dots \neq r_m$

then

$$g_0 = \frac{1}{2\pi j} \int \frac{G^*(z)}{z} dz = \left[ G(0) + \sum_{i=1}^m \frac{N(r_i)}{r_i M'(r_i)} \right] \quad (20)$$

and

$$g_n = \frac{1}{2\pi j} \int G^*(z) z^{n-1} dz = \sum_{i=1}^m \frac{N(r_i)}{M'(r_i)} r_i^{n-1} \equiv \sum_{i=1}^m \Phi(r_i) r_i^{n-1}, \quad n=1, 2, \dots \quad (21)$$

where

$$\Phi(r_i) = \frac{N(r_i)}{M'(r_i)}$$

Denoting the input pulse series of the controlled element  $G^*(z)$  by  $c_i^*(t) = \sum_{i=0}^{\infty} c_i \delta(t-iT)$ , and the sampled output due to only the single input pulse  $c_i z^{-i}$  by  $O^i(t)$ , then,

$$Z[O^i(t)] = G^*(z) c_i z^{-i} \quad (22)$$

For simplicity, the case of  $l=2$  is considered where the gain constants  $\xi_i$  in Figure 6 take the two values  $\xi_1$  and  $\xi_2$  alternately with the period  $T$ . The sampled error signal  $\epsilon_n$  at  $t = nT$  is as follows;

$$n=0: \quad \epsilon_0 = \frac{1}{(1+\xi_1 g_0)} \equiv A_1, \quad c_0 = \xi_1 \epsilon_0 \quad (23)$$

$$\therefore O^0(t) = [g_0 \delta(t) + \sum \Phi \delta(t-T) + \sum \Phi \cdot r_i \delta(t-2T) \\ + \sum \Phi \cdot r_i^2 \delta(t-3T) + \dots] \xi_1 A_1$$

$n=1$ :

$$\varepsilon_1 = -\frac{\Sigma \Phi \xi_1 A_1}{(1 + \xi_2 g_0)} \equiv -\Sigma \Phi \xi_1 A_1 A_2 \equiv -[\Phi][\xi_1 A_1] A_2 \quad (24)$$

where the symbol  $[\ ]$  denotes a matrix, and

$$[\Phi] \equiv [\Phi(r_1), \Phi(r_2), \dots, \Phi(r_m)] \quad (25)$$

$$[\xi_1 A_1] \equiv [1, 1, \dots, 1]_t \xi_1 A_1 \quad [\ ]_t: \text{transposed matrix}$$

$$\therefore O^{1*}(t) = [g_0 \delta(t-T) + \Sigma \Phi \delta(t-2T) + \Sigma \Phi r_i \delta(t-3T) + \dots] \times [-\Sigma \Phi \xi_1 A_1 \xi_2 A_2]$$

$n=2$ :

$$\begin{aligned} \varepsilon_2 &= [-\Sigma \Phi r_i \xi_1 A_1 + \Sigma \Phi \xi_1 A_1 \Sigma \Phi \xi_2 A_2] A_1 \\ &\equiv [\Phi][[\Phi \xi_2 A_2] - [r]] [\xi_1 A_1] A_1 \end{aligned} \quad (26)$$

where

$$[\Phi \xi_2 A_2] = \begin{bmatrix} \Phi(r_1) & \Phi(r_2) & \dots & \Phi(r_m) \\ \vdots & \vdots & & \vdots \\ \Phi(r_1) & \Phi(r_2) & \dots & \Phi(r_m) \end{bmatrix} \xi_2 A_2, \quad [r] = \begin{bmatrix} r_1 & 0 & \dots & 0 \\ 0 & r_2 & & \vdots \\ \vdots & \vdots & \ddots & 0 \\ 0 & \dots & 0 & r_m \end{bmatrix} \quad (27)$$

$$\begin{aligned} \therefore O^{2*}(t) &= [g_0 \delta(t-2T) + \Sigma \Phi \delta(t-3T) + \Sigma \Phi r_i \delta(t-4T) + \dots] \\ &\times [-\Sigma \Phi r_i \xi_1 A_1 + \Sigma \Phi \xi_1 A_1 \Sigma \Phi \xi_2 A_2] \xi_1 A_1 \end{aligned}$$

$n=3$ :

$$\begin{aligned} \varepsilon_3 &= [-\Sigma \Phi r_i^2 + \Sigma \Phi r_i \Sigma \Phi \xi_2 A_2 + \Sigma \Phi r_i \Sigma \Phi \xi_1 A_1 \\ &\quad - (\Sigma \Phi)^3 \xi_1 A_1 \xi_2 A_2] \xi_1 A_1 A_2 \\ &= [\Phi] [-[r]^2 + [r][\Phi \xi_1 A_1] + [r][\Phi \xi_2 A_2] \\ &\quad - [\Phi \xi_1 A_1][\Phi \xi_2 A_2]] [\xi_1 A_1] A_2 \\ &= -[\Phi][[\Phi \xi_1 A_1] - [r]][[\Phi \xi_2 A_2] - [r]][\xi_1 A_1] A_2 \end{aligned} \quad (28)^*$$

where

$$[\Phi \xi_1 A_1] = \begin{bmatrix} \Phi(r_1) & \Phi(r_2) & \dots & \Phi(r_m) \\ \vdots & \vdots & & \vdots \\ \Phi(r_1) & \Phi(r_2) & \dots & \Phi(r_m) \end{bmatrix} \xi_1 A_1 \quad (29)$$

In general,

$$\varepsilon_{2n+1} = -[\Phi] \{ [[\Phi \xi_1 A_1] - [r]][[\Phi \xi_2 A_2] - [r]] \}^n [\xi_1 A_1] A_2$$

$$\begin{aligned} \varepsilon_{2n} &= [\Phi][[\Phi \xi_2 A_2] - [r]] \\ &\times \{ [[\Phi \xi_1 A_1] - [r]][[\Phi \xi_2 A_2] - [r]] \}^{n-1} [\xi_1 A_1] A_1 \end{aligned} \quad (30)^{**}$$

Therefore, by using the characteristic roots of the matrix

$$[[\Phi \xi_1 A_1] - [r]][[\Phi \xi_2 A_2] - [r]]^\dagger$$

\*  $[\Phi \xi_i A_i][r] \neq [r][\Phi \xi_i A_i]$ , but  $[\Phi] \Phi \xi_i A_i [r] [\xi_i A_i] = [\Phi][r][\Phi \xi_i A_i][\xi_i A_i]$ .

\*\* In the equations,  $\xi_1 \neq 0$  is assumed. If  $\xi_1 = 0$ , the system is never disturbed, and then  $\xi_2$  must be substituted for  $\xi_1$ . If  $\xi_1 = 0$  and  $\xi_2 = 0$  at the same time, then all input pulses  $x_1, x_2$  are saturated and the stability of the oscillation depends on whether  $G^*(z) D_F^*(z)$ , the open-loop transfer function in the original system, is stable or not.

†  $[[\Phi \xi_2 A_2] - [r]][[\Phi \xi_1 A_1] - [r]]$  may be also used.

that is, the roots  $R_i$  of the equation

$$\det. \{ R[U] - [[\Phi \xi_1 A_1] - [r]][[\Phi \xi_2 A_2] - [r]] \} = 0 \quad (31)$$

where  $[U]$  denotes a unit matrix, the stability of the system in Figure 6 is determined as follows<sup>1</sup>. If all absolute values of the characteristic roots are smaller than unity, the system is stable and, if not, it is unstable. It is also possible to apply the Hurwitz stability criterion to this problem by introducing the following bi-linear transformation into eqn (31).

$$R = \frac{(\lambda + 1)}{(\lambda - 1)} \quad (32)$$

In general, the stability criterion of the equilibrium state of period  $IT$  is given by the absolute values of the roots of

$$\det. \{ R[U] - \prod_{i=1}^l [[\Phi \xi_i A_i] - [r]] \} = 0 \quad (33)$$

### Examples<sup>2</sup>

**Example 1**—The system shown in Figure 7(a) is a finite-settling-time system for both a step input  $r(t)$  and a step disturbance  $d(t)$ . A numerical example of the responses, when a single impulse of the magnitude  $\gamma_0$  is applied to the system, is shown in Figure 7(b). From the diagram, it is seen that the oscillation of period  $IT = 2T = 0.5$  is caused thereby and that it builds up when  $\gamma_0 > 0.4831 \Delta$ , but vanishes when  $\gamma_0 < 0.4831 \Delta$ . It is presumed that the unstable equilibrium state of the oscillation of period  $IT = 2T$  may be obtained when  $r_0 = 0.4813 \Delta$ .

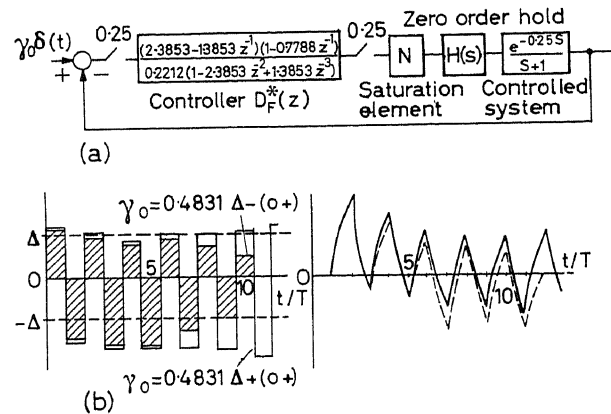


Figure 7. (a) A finite-settling-time response system (1); (b) Transient response to an impulse input  $\gamma(t) = \gamma_0 \delta(t)$

Such phenomena are analogous to 'the hard self-excitation' in continuous systems<sup>3</sup>.

In this system,

$$B_1 = 1.3853, A_1 = -1.3853, B_2 = -1.3853, A_2 = 2.3853 \quad (34)$$

and the following two cases are considered as the possible equilibrium states:

$$(1) \quad x_1 \geq \Delta, x_2 \leq -\Delta \quad (\therefore \alpha_1 x_1 = \Delta, \alpha_2 x_2 = -\Delta) \quad (35)$$

In this case, eqn (5) becomes

$$\begin{aligned} B_1 x_1 + B_2 x_2 &= -(A_1 - A_2) \Delta, \\ B_2 x_1 + B_1 x_2 &= (A_1 - A_2) \Delta \end{aligned} \quad (36)$$

$$\therefore x_1 - x_2 = 2.72 \Delta \quad (37)$$

Here, two equations (36) being linearly dependent, each value of  $x_1$  and  $x_2$  cannot be uniquely determined by eqn (36). As illustrated in Figure 8, the solutions which satisfy eqns (35) and (36) exist infinitely in the intervals of

$$\Delta \leq x_1 \leq 1.72 \Delta, \quad -1.72 \Delta \leq x_2 \leq -\Delta \quad (38)$$

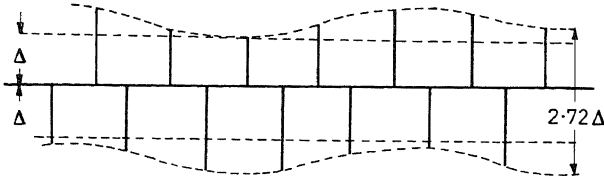


Figure 8. Periodic equilibrium state

This suggests that the mean value of the controller output should be given as another complementary condition, because the controller  $D_F^*(z)$  has a pole at  $z = 1$  and the magnitudes of both pulses  $x_1$  and  $x_2$  are limited to constant values  $\pm \Delta$  by the saturation element independently of the mean value of the controller output, and each value of  $x_1$  and  $x_2$  can be uniquely determined when the mean value of the controller output is given.

$$(2) \quad x_1 \geq \Delta, |x_2| \leq \Delta (\therefore \alpha_1 x_1 = \Delta, \alpha_2 = 1) \quad (39)$$

$$B_1 x_1 + (B_2 + A_2) x_2 = -A_1 \Delta$$

$$B_2 x_1 + (B_1 + A_1) x_2 = -A_2 \Delta$$

$$\therefore x_1 = 1.72 \Delta, (\alpha_1 = 0.585), x_2 = -\Delta, (\alpha_2 = 1) \quad (40)$$

The above solutions are limiting values of eqn (38).

Consider the stability of these equilibrium states. In this system

$$r_1 = 1, r_2 = 0.779, r_3 = -1.779, \Delta_1 = \Delta_2 = 1 \quad (41)$$

$$\phi(r_1) = 1.628, \phi(r_2) = -0.836, \phi(r_3) = -0.792$$

and

$$\xi_1 = 0 \text{ and } \xi_2 = 0 \text{ for } x_2 \neq \Delta, \text{ or } 0 \leq \xi_2 \leq 1 \text{ for } x_2 = -\Delta \quad (42)$$

$$\begin{aligned} \therefore \det. \{R[U] - [\Phi \xi_1 A_1] - [r]\} [\Phi \xi_2 A_2] - [r]\} \\ = R^3 - (4.772 - 1.395 \xi_2) R^2 + (5.693 - 3.328 \xi_2 - 0.313 \xi_2^2) R \\ + (-1.921 + 1.123 \xi_2 + 0.313 \xi_2^2 - 0.299 \xi_2^3) = 0 \end{aligned}$$

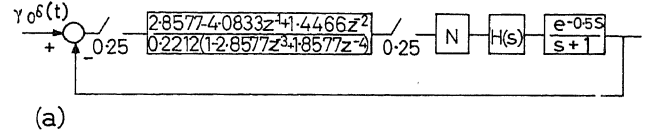
$$\begin{aligned} \therefore -(0.810 + 0.299 \xi_2^2) \xi_2 \lambda^3 \\ + (1.702 + 1.354 \xi_2 - 0.626 \xi_2^2 + 0.897 \xi_2^3) \lambda^2 \\ + (-3.684 + 5.302 \xi_2 + 1.252 \xi_2^2 - 0.897 \xi_2^3) \lambda \\ + (13.39 - 5.846 \xi_2 - 0.626 \xi_2^2 + 0.299 \xi_2^3) = 0 \end{aligned} \quad (43)^*$$

By considering the signs of coefficients of  $\lambda^3$ ,  $\lambda^2$  and  $\lambda$  it can be concluded that eqn (43) has at least one unstable root for  $0 \leq \xi_2 \leq 1$  and, hence, that this equilibrium state is unstable.

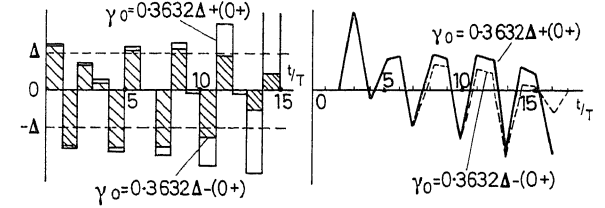
\* When the saturation characteristics are idealized as shown in Figure 4  $\xi_i$  changes at  $|x_i| = \Delta$  discontinuously, but  $\xi_i$  changes continuously from 0 to 1 for the practical element.

Example 2—Figure 9(a) shows another finite-settling-time response system with the delay time  $2T = 0.5$ . From the numerical examples illustrated in Figure 9(b) the presence of the unstable equilibrium state of period  $3T = 0.75$  may be estimated. By the same procedure as in Example 1, the following unstable periodic equilibrium state of period  $3T$  can be confirmed,

$$x_1 = 1.55 \Delta > \Delta, \quad x_2 = -\Delta, \quad x_3 = 0 \quad (44)$$



(a)



(b)

Figure 9. (a) A finite-settling-time response system (2); (b) transient response to an impulse input  $\gamma(t) = \gamma_0 \delta(t)$

### Describing Function for Non-linear Sampled-data Systems

For the analysis of steady-state phenomena in non-linear sampled-data systems, a so-called describing function can be introduced as a useful tool as in continuous systems. However, the most important subject, in this case, is the method by which the sampler must be treated.

#### The First Method

A sustained oscillation of frequency  $\omega$  in the non-linear sampled-data control system is assumed as illustrated in Figure 10.

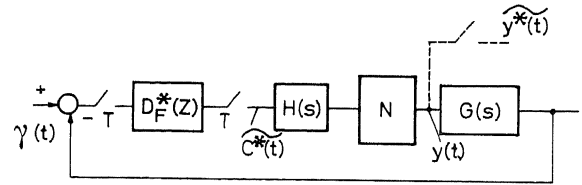


Figure 10. Non-linear sampled-data system

Now assuming that  $\omega \ll \omega_s$ , where  $\omega_s = 2\pi/T$  is the sampling frequency, and that the sampled controller output  $\tilde{c}^*(t)$  is

$$\tilde{c}^*(t) = c \sum_{i=-\infty}^{\infty} \sin(\omega i T + \varphi_c) \delta(t - iT) \quad (45)$$

then the continuous output of the non-linear element  $\tilde{y}(t)$  and its sampled pulse series  $\tilde{y}^*(t)$  are approximated by

$$\begin{aligned} \tilde{y}(t) &= N(x) x \sin(\omega t + \varphi_c + \varphi_h), \\ x &= |H(j\omega)| \omega_s C |2\pi, \quad \varphi_h = \angle H(j\omega) \end{aligned} \quad (46)$$

$$\tilde{y}^*(t) = N(c) c \sin(\omega i T + \varphi_c) \delta(t - iT) \quad (47)$$



where  $N(x)$  and  $N(c)$  are the describing functions of the saturation element and  $j$  is  $\sqrt{-1}$ .  $N(c)$  is a describing function for the sinusoidal envelope of controller output pulse series, and  $N(x)$  is that for the fundamental sinusoid of staircase wave of the hold-circuit output as illustrated in Figure 11. Corresponding to eqns (46) and (47), the two kinds of balancing equations for self-sustained oscillations are given, respectively, as follows;

$$1 + N(x) \phi_1(\omega) = 0, \quad \phi_1(\omega) = \frac{\omega_s}{2\pi} D_F^*(e^{j2\pi \frac{\omega}{\omega_s}}) H(j\omega) G(j\omega) \quad (48)$$

$$1 + N(c) \phi_1^*(\omega) = 0, \quad \phi_1^*(\omega) = D_F^*(e^{j2\pi \frac{\omega}{\omega_s}}) H G^*(e^{j2\pi \frac{\omega}{\omega_s}}) \quad (49)$$

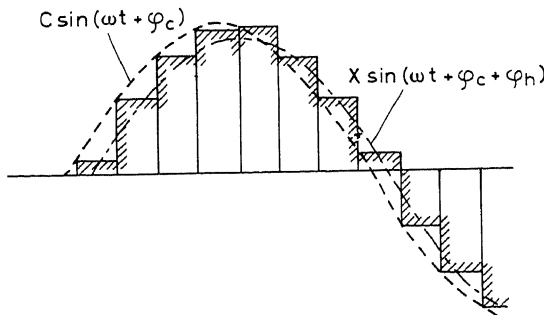


Figure 11. Hold circuit output waveform

The phase angle  $\varphi_c$  is not contained in eqns (48) and (49); in other words, the equilibrium states are determined independently of the phase angle. Consequently, the accuracy of this method decreases as  $\omega$  tends to  $\omega_s$  and the sampled signals such as  $\tilde{c}^*(t)$  and  $\tilde{y}^*(t)$  cannot be determined unless the other additional conditions concerning the phase angle are given.

The comparison of the utility of the above equations is given in the following examples.

**Example 3**—In Figure 10, supposing that the non-linear element  $N$  is the saturation element as shown in Figure 4, and that the transfer functions  $G(s)$ ,  $H(s)$  and  $D_F^*(z)$  are given by

$$G(s) = \frac{1}{s(s+1)}, \quad H(s) = \frac{1-e^{-sT}}{s}, \quad T=0.25$$

$$D_F^*(z) = \frac{68.509 - 74.104z^{-1} + 23.678z^{-2}}{1 - 0.19447z^{-1} - 0.86553z^{-2}} \quad (50)$$

then the system is a finite-settling-time system. Figure 12 shows a numerical example of the controller output, when a step function of magnitude  $\gamma$  is applied as the input  $\gamma(t)$ . The 'hard self-excitation' takes place also in this system, as illustrated in Figure 12. Figure 13 shows Nyquist diagrams by eqns (48) and (49). In the figure, the cross point of the frequency vector locus  $\phi_1(\omega)$  and the amplitude locus  $-1/N(x)$  coincides with that of  $\phi_1^*(\omega)$  and  $-1/N(c)$ . The cross point is an unstable equilibrium point. The value of the point is  $\omega T = 0.57$ , (that is  $\omega \cong \omega_s/11$ ) and  $C = x = -4.45$ . The broken lines in Figure 12 indicate the equilibrium amplitude  $C = x = 4.45$ . These values of  $\omega$  and  $C$  (or  $x$ ) can fairly explain the phenomena shown in Figure 12.

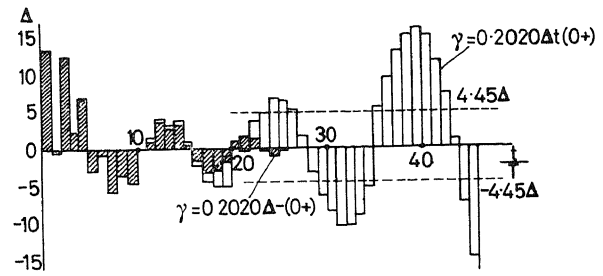


Figure 12. Transient response

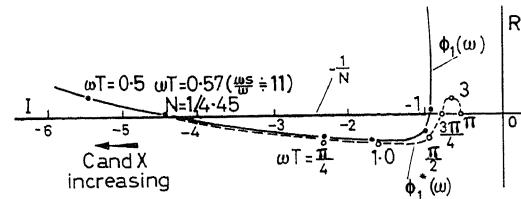


Figure 13. Frequency and amplitude vector loci

**Example 4**—Figures 14 and 15 show Nyquist diagrams with the amplitude locus of  $-1/N$  for the systems illustrated in Examples 1 and 2 respectively. In Figure 14, the vector locus of  $\phi_1(\omega)$  does not cross the amplitude locus and so the presence of any equilibrium oscillation cannot be expected. But both vector loci of  $\phi_1^*(\omega)$  in Figures 14 and 15 have the unstable cross point coinciding with the amplitude locus; hence, in this case, the presence of the unstable equilibrium state can be expected. However, the numerical values of  $\omega$  and  $C$  given by the cross points differ from the accurate values given earlier.

#### The Second Method<sup>4</sup>

A describing function for  $\omega \cong \omega_s$  is introduced by considering the phase angle of the oscillation at the sampling instants. To simplify the analysis, the following restrictions should be applied.

(1)  $G(s)$  or  $D_F^*(z)$  is assumed to include at least an integrative pole. From this assumption, it can be concluded that the mean

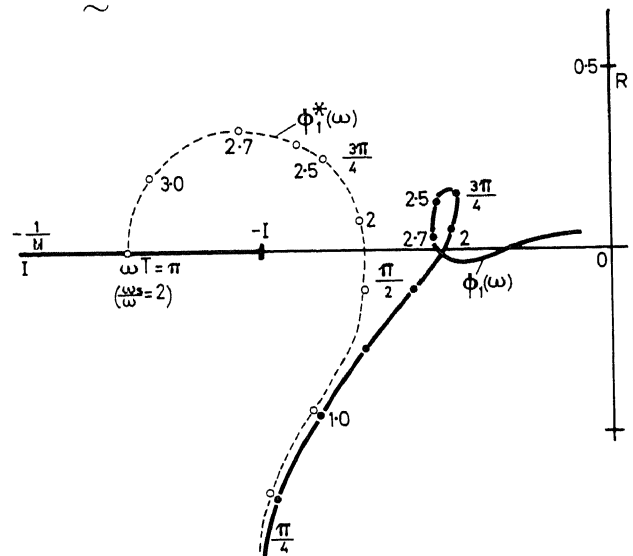


Figure 14. Frequency and amplitude vector loci

value of  $y(t)$  over a period must be zero for the continuation of the periodic phenomenon, and that the ratio  $\omega_s/\omega$  is an integer number such as 2, 3, 4 etc., because the existence of a zero-order hold makes the sign changes of  $\tilde{y}(t)$  take place only at sampling instants as illustrated in Figure 16.

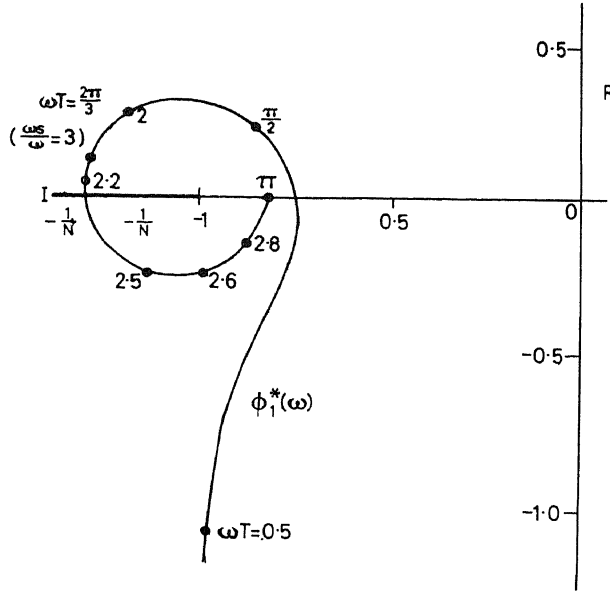


Figure 15. Frequency and amplitude vector loci

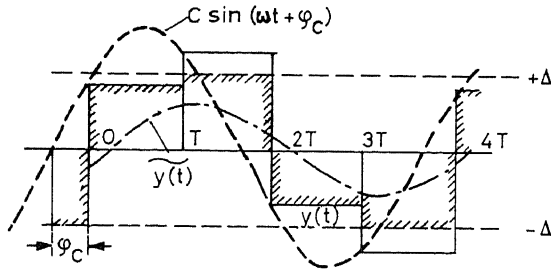


Figure 16. Phase relation between  $c(t)$ ,  $y(t)$  and  $\tilde{y}(t)$

(2) The periodic phenomenon is assumed to be a sinusoidal oscillation of frequency  $\omega$ . When  $\tilde{c}^*(t)$  passes through the zero order hold and the saturation element, the output  $y(t)$  becomes a staircase wave as represented by hatching in Figure 16. Denoting the fundamental component of the output  $y(t)$  by  $\tilde{y}(t)$ ,

$$\tilde{y}(t) = \text{Im} (N_{\omega_s/\omega}(c, \varphi_c) c e^{j(\omega t + \varphi_c)}) \quad (51)$$

where  $N_{\omega_s/\omega}(c, \varphi_c)$  is a describing function of the saturation element combined with a zero order hold. It is a function not only of  $c$  and  $\varphi_c$  but also the value of  $\omega_s/\omega$  and the number of saturated pulses in a period. The balancing relation in this case is

$$1 + N_{\omega_s/\omega}(c, \varphi_c) \phi_2(\omega) = 0, \phi_2(\omega) \equiv D_F^*(e^{j2\pi\frac{\omega}{\omega_s}}) G(j\omega) \quad (52)$$

*The Case  $\omega_s/\omega$  is an Even Number*—When  $\omega_s/\omega$  is an even number, the output of the saturation element has no d.c. component independently of the value of  $\varphi_c$  and practically, by considering only the region of  $-\omega T/2 \leq \varphi_c \leq \omega T/2$ , the phenomena can be estimated.

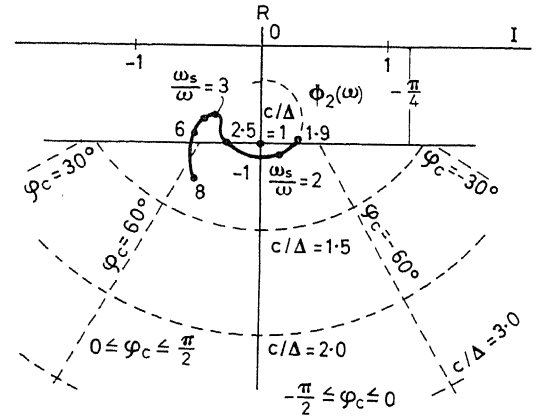
(1)  $\omega_s/\omega = 2$ —In this case evidently

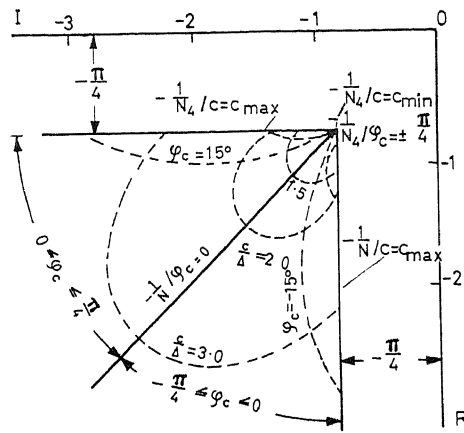
$$0 \leq \Delta \leq c \sin \varphi_c : C_{\min} = \frac{\Delta}{\sin \varphi_c}, \quad C_{\max} = \infty \quad (53)$$

$$-\frac{1}{N_2(c, \varphi_c)} = -\frac{\pi c e^{j\varphi_c}}{4\Delta}, \quad -\frac{1}{N_2|_{C_{\min}}} = -\frac{\pi e^{j\varphi_c}}{4 \sin \varphi_c} \text{ for } 0 \leq \varphi_c \leq \frac{\pi}{2}$$

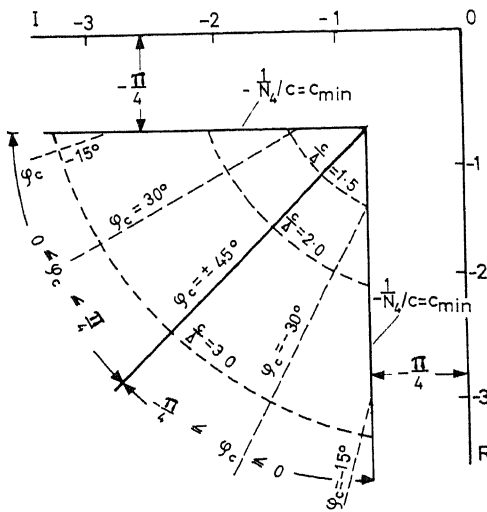
$$-\frac{1}{N_2(c, \varphi_c)} = \frac{\pi c e^{j\varphi_c}}{4\Delta}, \quad -\frac{1}{N_2|_{C_{\min}}} = \frac{\pi e^{j\varphi_c}}{4 \sin \varphi_c} \text{ for } -\frac{\pi}{2} \leq \varphi_c \leq 0 \quad (54)$$

Figure 17 is a vector representation of eqn (54). The frequency locus  $\phi_2(\omega)$  plotted in the diagram is the same system illustrated in Example 1. From this, the point corresponding to  $\omega_s/\omega = 2$  falls inside the  $-1/N_2$  domain and its coordinates give the periodic equilibrium state of period  $2T$ .





(a)



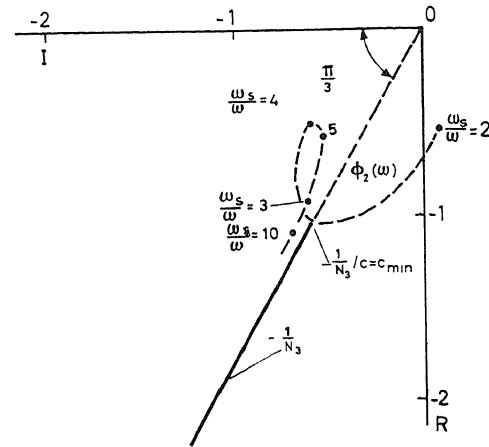
(b)

Figure 18. Vector representation of  $-1/N_4$ :  
(a) Mode 1; (b) Mode 2

Evidently the vector locus of eqn (58) is a straight line as indicated in Figure 19, and the frequency vector locus  $\phi_2(\omega)$  in the diagram is the same as that in Example 2. The point corresponding to  $\omega_s/\omega = 3$  is very near to the locus of  $-1/N_3$ , but not on it. This discrepancy is due to restriction (1). That is, the actual waveform is quite different from the sinusoidal wave as indicated in Figure 9(b).

### Conclusion

The analysis described first does not include any assumption and is applicable to non-sinusoidal oscillations and to the

Figure 19. Vector representation of  $-1/N_3$ 

phenomena in higher order systems. But this analysis is restricted to the phenomena of period  $lT$  where  $l$  is an integer; and when  $l$  is increased, the analysis becomes extremely difficult. And moreover, to estimate the presence of the oscillation, the analysis must be carried out for various values of  $l$ .

By the first describing function methods proposed, the periodic equilibrium states and their stabilities are very easily analysed. But the accuracy of the results decreases rapidly as the frequency of oscillation approaches the sampling frequency, because, in these methods, the waveform is assumed to be sinusoidal and the phase angle of oscillation is not taken into account.

In the second method, the phase angle of the oscillation at the sampling instants is taken into consideration, and this method is applicable to the case where both frequencies approach closely. But in this method, the assumption that the system includes at least one integrative pole is required in addition to the usual assumptions in a describing function method. Thus the stability of the equilibrium state cannot be dealt with by this method.

### References

- 1 FRAZER, R. A., DUNCAN, F., and COLLAR *Elementary Matrices and Some Applications to Dynamics and Differential Equations*. 1938. Cambridge
- 2 KONDO, B., IWAI, S., and SOGA, M. Finite settling-time response in sampled-data control systems with saturation. *Proc. 1st IFAC Conf.* p. 305. 1961. London; Butterworths
- 3 TRUXAL, J. G. *Control System Synthesis*, p. 641. 1955. New York; McGraw-Hill
- 4 CHOW, C. K. Contactor servomechanisms employing sampled-data. *Trans. Amer. Inst. elect. Engrs* 73 (1954) 55

### DISCUSSION

J. PIKIELNY, *Przemysłowy Instytut Telekomunikacji, Warszawa, Białowiecka 19/47, Poland*

Describing function methods have been used for analysis of linear and non-linear sampled-data systems for some time. This method is also

mentioned in your paper, and I therefore ask if you have made any attempts, either theoretical or empirical, to estimate the accuracy of this method and what results were obtained. It is very interesting to compare the results obtained by first and second (Chow's) describing function methods.

B. KONDO, *in reply*

Thank you for the discussion of our paper. The first method is clearly less accurate than the second one. The main reason might be the lack of the phase difference between the sustained oscillation and the samples in describing function  $N(c)$  or  $N(x)$ .

For example, consider the extreme case where the saturation level is very small compared with the amplitude of oscillation. In this case, it can be seen that the output is always the same when any sinusoidal input with an arbitrary phase angle between  $-\pi/l$  and  $\pi/l$  is applied. This means that the phase angle of describing function varies with the phase difference between the samples and the oscillation (this is very clear in Chow's method). Thus, roughly speaking, it can be seen that the maximum phase error is  $\pm \pi/l$  for the saturation element. Similarly, the gain error of describing function can be considered.

However, the above conclusion is unfortunately limited to the curve where  $\omega/\omega_s$  is an integer, because the describing function in the strict meaning cannot be defined in our system<sup>1</sup>. This problem remains to be solved in the future.

#### Reference

- <sup>1</sup> KALMAN, R. E. Non-linear mechanics Symp., Polytechnic Inst. Brooklyn U.S.A. (1956)

E. PAVLIK, *Siemens and Halske, Karlsruhe, Lassallestr. 9, W. Germany*

The analysis is restricted to the phenomena of period  $lT$  where  $l$  is an integer.

In Example 1 it was empirically stated that there was such an integer  $lT$ . Must this integer *always* be stated empirically or does an analytical method exist?

B. KONDO, *in reply*

We are sorry to say that, as far as we know, the method of determining directly the possible value of  $l$  has not yet been found. In our method, assuming the value of  $l$ , eqn (5) can be solved. If a reasonable solution can be obtained, then it can be concluded that the oscillation with the assumed period exists.

We would add, however, that Professor Jury and Dr. Nishimura have succeeded in formulating the equation by which the possible maximum period of the oscillation can be determined under some restriction for PWM systems.

M. HAMZA, *ETH, Zurich, Switzerland*

Could Professor Kondo please indicate whether his method can be modified to cover systems with hysteresis? If not, which method does he suggest?

B. KONDO, *in reply*

Equation (5) in our paper can also be applied to the system with hysteresis. In this case, we need the assumption for formulation of eqn (5) that the operating point corresponding to a given  $x_i$  is on either the upper region or the lower region of the hysteresis. Under this assumption, we can formulate the equation. However, it is necessary to examine if the obtained solution of eqn (5) satisfies the above-mentioned assumption.

A. LEPSCHY and A. RUBERTI, *I.S.P.T. Viale Trastevere 189, Roma, Italy*

We intervene in the use of the describing function for the analysis of limit cycles, neglecting the first method which, as the authors themselves affirm and which Tsytkin<sup>1</sup> has shown, holds only if the rate  $\omega_s/\omega$  is high. As far as the second method is concerned, we feel that it will be useful to communicate our studies with an analogous

procedure in a more general case<sup>2</sup>. Taking into account the fact that the method applies only if the block which contains the sampler and the non-linear element is represented by one describing function only<sup>3-6</sup>, we proposed to elaborate a simple procedure for the analysis of a generic instantaneous piecewise non-linearity.

We have therefore considered the sum of a number of dead zones or of limiting characteristics instead of the actual non-linear characteristic; then we have studied the approximation limits within which it is allowed to sum the describing functions of the elements (formed by a sampler, a holder and a dead zone or a saturation) the parallel of which is substituted for the block formed by the sampler, the holder and the actual non-linear element.

We have preliminarily investigated the case in which the non-linear element is a dead zone or a saturation. (The latter case is the same as that studied by Messrs. Kondo and Iwai.) In this analysis we thought it necessary to take into account the effects of the continuous component at the input of the non-linear element (the continuous component at the input is there if the rate  $\omega_s/\omega$  is even and  $G(s)$  has a pole at the origin, which prevents the non-linear element from having a continuous component at its output). The study of the amount of the continuous component<sup>7</sup> and of its effects, is important with regard to the above-mentioned decomposition of the actual non-linear element. In our paper we present the normalized diagrams of the real and imaginary parts of the describing function of the block formed by sampler, holder and dead zone. Each diagram is calculated for one value of  $\omega_s/\omega$  (from 2 to 10) and shows how the describing functions depend on the amplitude of the input signal and on its phase with respect to the sampling impulses. The diagrams give the real and the imaginary parts of the describing functions in order to facilitate their composition.

#### References

- <sup>1</sup> TSYTKIN, YA. Z. Elements of the theory of computing automatic systems. *Proc. 1st I.F.A.C. Congr.*, Moscow. 1961. London; Butterworths
- <sup>2</sup> LEPSCHY, A. and RUBERTI, A. The describing function for the study of sampled-data control systems with a piecewise non-linearity. *Alta Frequenza* (English issue), XXXII, 5, March 1963, p. 357 (85E)
- <sup>3</sup> CHOW, C. K. Contactor servomechanisms employing sampled data. *Trans. Amer. Inst. elect. Engrs.*, 73, II (1954) 51
- <sup>4</sup> RUSSEL, F. A. Discussion of paper 'Contacter servomechanisms employing sampled-data'. *Trans. Amer. Inst. elect. Engrs.*, 73, II (1954)
- <sup>5</sup> KAZAKAN, V. P. Investigation of the dynamic relay pulse systems of automatic control (in Russian). *Auto. Telemekh.*, 18, 8 (August 1957)
- <sup>6</sup> JURY, H. I. Recent advances in the field of sampled data and digital control systems. *Proc. 1st I.F.A.C. Congr.*, Moscow. 1961. London; Butterworths
- <sup>7</sup> LEPSCHY, A. and RUBERTI, A. La funzione descrittiva per lo studio dei sistemi a segnali campionati comprendenti non linearità a tratti (Appendice). *Alta Frequenza*, XXXII, 1963

B. KONDO, *in reply*

Thank you for your very valuable comments on our paper. The results achieved with your investigations of similar problems are very useful, and it is hoped that they will be of assistance in our study in the future.

G. SENOUILLET, R. BOUDAREL and P. GUICHET, *Centre d'Etude et de Recherche en Automatisme, 32, Boulevard Victor, Paris 15<sup>e</sup> France*

The research of the conditions of existence of periodic limit cycles is greatly simplified by the use of a sequence vector which provides a unifying concept to the methods proposed by different authors.

A sampled periodic oscillation is represented by a vector in an orthonormal space the coordinates of which are the values of the successive samples. The number of coordinates of the vector is equal to the number of samples included during one period of oscillation.

Then, considering a non-linear sampled data system consisting of a number of linear elements separated by instantaneous non-linearities, and assuming that there exists a limit cycle of period  $NT$ , the corresponding oscillation can be represented by a sequence vector  $\mathbf{E}$  defined numerically or parametrically at a given point of the loop. Opening the loop at this point it is possible to compute the transmission of the sequence vector  $\mathbf{E}$  throughout the loop. The assumed oscillation will exist when the output sequence vector  $\mathbf{S}$  will be identical to the input sequence vector  $\mathbf{E}$ . This yields the formulation of sufficient and necessary conditions of existence of the assumed oscillation.

The practical problem arises when the computation of the transmission of the sequence vector is to be performed. For this, the sequence vector can be given several representations, depending upon the choice of the bases of the orthonormal space.

One type of representation is transient, another type is harmonic.

The time representation may be direct, i.e. the successive coordinates are the successive samples of the oscillation; or inverse, i.e. the successive coordinates being taken backward of the successive samples, coordinate zero being the value of the signal at the sampling instant chosen as origin, the coordinate number one being the value of the signal at the former sampling time, etc. The harmonic representations are defined in relation to the direct time representation.

One can consider first a base, the vectors of which are

$$\mu_i^k = \cos\left(\frac{2\pi mk}{N} + \phi\right), \quad m \text{ integer } \geq 0$$

where  $\mu_i^k$  is the  $k$ th coordinate of vector  $\mu_i$  with respect to the direct transient base. Then a periodic oscillation of period  $N$  can be represented by the limited expansion

$$s(k) = a_0 + a_1 \cos \frac{2\pi k}{N} + a_2 \sin \frac{2\pi k}{N} + \dots$$

where the  $a_i$ 's are the coordinates of the sequence vector with respect to the new base.

One can also consider the direct transient space as Hermitian and choose a new base the vectors of which are

$$\lambda_i^k = \exp\left(\frac{2\pi m}{N} kj\right)$$

which yields a closely-related representation of the sequence vector.

Depending upon the chosen representation, several methods can be developed for computing the transmission of the sequence vector throughout a loop consisting of linear sampled elements separated by instantaneous non-linearities.

First it is to be noted that only the time representation yields a straightforward expression of the action of the non-linearities since it appears successively on each sample, that is, on each coordinate.

If the non-linearity is given by

$$s = f(e)$$

where  $e$  is the input of the non-linearity and  $s$  is the output of the non-linearity, the  $k$ th coordinate of the output sequence vector  $\mathbf{S}$  will be related to the  $k$ th coordinate of the input sequence vector  $\mathbf{E}$  by:

$$s_k = f(e_k)$$

In the other representations, where coordinates are not successive samples, the effect of the non-linearity cannot be separated for each coordinate. It is therefore convenient to come back to the time representation by performing the coordinate transformation.

The transmission of the sequence vector through a linear sampled element can be computed in time representations and in harmonic representations.

#### Transmission of the Sequence Vector in Time Representation

The output of the linear element is constituted by an oscillation of the same period as the input. Therefore, the transformation in the sequence vector space yielding the output vector  $\mathbf{S}$ , starting from the input vector  $\mathbf{E}$ , is a linear transformation.

$$\mathbf{S} = \mathbf{W}\mathbf{E}$$

where  $\mathbf{W}$  is a matrix.

The choice of the inverse time representation yields a particularly simple structure of the transmission matrix which has only real elements.

Naturally the presence in the linear element of a pole at 1 makes that the transformation degenerates and special treatment must be reserved to the integration factors which would be separated from the rest of the element.

Such a matrix has interesting properties<sup>1</sup>, for example, (a) They are associative; (b) They are commutable; (c) They are invariant in time translation for stationary filters and therefore are circulant, i.e. invariant in circular permutations of lines and columns; and (d) Noting  $\mathbf{P}$  as the circular permutation matrix, the sampled transfer function  $\mathbf{Z}^{-1}$  corresponds to the transmission matrix  $\mathbf{P}$  and therefore the filter  $\mathbf{F}(\mathbf{Z}^{-1}) = \sum f_i \mathbf{Z}^{-i}$  has for transmission matrix

$$\mathbf{W} = \sum f_i \mathbf{P}^i = \mathbf{F}(\mathbf{P})$$

Matrices  $\mathbf{W}$  can be computed by using  $\mathbf{Z}$  transform techniques and the use of the skip sampling theorem<sup>2</sup> greatly simplifies the evaluation of the integrals involved. The Frazer-Duncan-Collar formula can be used in order to compute directly

$$\mathbf{W} = \mathbf{F}(\mathbf{P})$$

Matrix techniques provide a straightforward method to compute the transmission matrix  $\mathbf{W}^{1-3}$ . The transfer function being written under the form of parallel or cascaded expansion, this method yields the transmission matrix  $\mathbf{W}$  for the elementary elements

$$\frac{b_i}{1 - C_i \mathbf{Z}} - 1$$

The general transmission matrix is the combination of the elementary transmission matrices.

#### Transmission of the Sequence Vector in Harmonic Representation

The representation in an Hermitian space yields a diagonal matrix the terms of which are complex numbers<sup>4, 5</sup>, while the representation in an Euclidian space yields a real anti-symmetric matrix<sup>6</sup>.

From this last representation, if the linear element is a low-pass element, it is sufficient to limit the expansion of the input and output sequence vectors to the two first coordinates. This constitutes the first harmonic approximation which was the first developed for the analysis of non-linear sampled-data systems<sup>7, 8</sup>. Considerations upon the representations of the sequence vectors show that, when the first harmonic hypothesis is valid, the first harmonic approximation is identical to the sequence vector itself. Therefore, there is no necessity to compute any harmonic decomposition of any sequence vector, the time representation being perfectly adequate<sup>9</sup>.

In sampled-data systems, because of the folding of frequency caused by sampling, the domain of application of first harmonic method is more restricted than in continuous systems. Moreover, the first harmonic approximation, when valid, provides only a method for sinusoidal type oscillation and it is well known that non-linear sampled-data systems exhibit non-sinusoidal type oscillations.

In fact, rigorous methods do not necessitate computations more intricate than the first harmonic approximation. Moreover, the matrix techniques can easily be mechanized and therefore allow one to cope with high-order systems with any number of non-linearities, as is shown in Reference 1 in the particular case of coded non-linearities.

The preceding considerations will be developed at some length in a paper 'Analytical Study of Non-Linear Sampled Data Systems', by R. Boudarel, P. Guichet and G. Senouillet, to appear in *Progress in Automatic Control*, Vol. 3 (edited by MacMillan).

#### References

- <sup>1</sup> BOUDAREL, R. Stabilité des Asservissements Echantillonnés et Quantifiés. Mémorial de l'Artillerie Française. 4e trimestre (1962). No. 142
- <sup>2</sup> JURY, E. I. and NISHIMURA, T. On the periodic modes of oscillation in PWM feedback systems. *Amer. Soc. mech. Engrs.* (July 1961)
- <sup>3</sup> BOUDAREL, R., SENOUILLET, G. and GUICHET, P. Etude Analytique des Auto-Oscillations dans les Systèmes Echantillonnés Non-Linéaires. *Automatisme* (July 1963)
- <sup>4</sup> TORNG, H. On periodic modes of saturating discrete control systems. *J. Franklin Inst.* 274, No. 5 (1962)
- <sup>5</sup> TSYPKIN, YA. Z. Elements of theory of numerical automatic systems. *Proc. 1st I.F.A.C. Congr. Moscow.* 1961. London; Butterworths
- <sup>6</sup> MESERVE, W. and TORNG, H. Investigation of periodic modes of sampled data control systems containing a saturating element. *Trans. Amer. mech. Engrs.* (March 1961)
- <sup>7</sup> CHOW, C. K. Contactor servomechanisms employing sampled data. *Trans. Amer. Inst. mech. Engrs.* 73-11 (1954)
- <sup>8</sup> SENOUILLET, G. and GUICHET, P. Stabilité d'un Asservissement Echantillonné et Quantifié. *Onde Electrique* (July 1961)
- <sup>9</sup> GUICHET, P. Sur les Oscillations Périodiques Présentées par les Systèmes Echantillonnés Non-Linéaires. *Automatisme* (June 1963)

R. E. KALMAN, *Research Institute for Advanced Studies, 7212 Bellona Avenue, Baltimore 12, MD., U.S.A.*

Since this is a carefully written paper, it is especially unfortunate that the authors seem unaware of the extensive literature concerning this problem.

It is well known<sup>1</sup> that Chow's method can be expected to hold only under highly restrictive assumptions. Both Chow's method and the first method of the paper were thoroughly explored by Torng<sup>2</sup>. I am obstinately of the opinion that transform methods are all wrong for non-linear problems. The 'natural' mathematical approach is always in terms of state variables, the success of this approach is amply clear from the literature.

#### References

- <sup>1</sup> KALMAN, R. E. *Non-linear Mechanics, Symp.* Polytechnic Inst. Brooklyn. U.S.A. 1956
- <sup>2</sup> TORNG, J. and MESERVE, M. E. *Trans. Amer. Soc. mech. Engrs; J. Basic Eng.* (1961)

B. KONDO, *in reply*

We would not say that our paper was the first in which the describing function was applied for the analysis of non-linear sampled-data systems. The main purpose of the paper lies in the proposition of the strict method, the description of which covers the first two-thirds of our paper, and I would be grateful if readers will bear this in mind when reading our paper.

YA. Z. TSYPKIN, *Institute of Automation and Telemechanics, Moscow I-53, Kalachevskaya, 15-A, U.S.S.R.*

The interesting paper by Messrs. Kondo and Iwai gives precise and approximate methods of determining periodic behaviour in non-linear sampled-data systems. A number of studies have been made in this field in the Soviet Union, and have been published in the journal *Automat. Telemekh.* from 1960 to 1963, and in the proceedings of a symposium on the theory of non-linear oscillations, Kiev, 1961. In a number of cases the results obtained differed from those of Kondo and Iwai only in form.

I should like to note that for determining symmetrical non-periodic behaviour in non-linear sampled-data systems it is more convenient to use, in place of the simultaneous eqns (2) and (3), the system

$$\tilde{x}[n] = \sum_{v=0}^{N-1} A_{n-v}(N) \varphi(\tilde{x}[v]) \quad n=0, 1, 2, \dots, N \quad (1)$$

where

$$A_{n-v}(N) = \frac{1}{N} \sum_{\mu=0}^{N-1} W^* \left( j \frac{\mu\pi}{N} \right) e^{-j \frac{\mu\pi}{N} (n-v)} \quad (2)$$

where  $W^*(j\bar{\omega})$  is the frequency characteristic of the linear sampled-data side and  $N$  the half-wave of the oscillations.

This system can easily be solved by iteration, and  $\tilde{x}[u]$  thus determined.

The stability of forced periodic behaviour *im Grossen* can be studied on the basis of the approach described in our paper read at this Congress.

I should like to mention the very interesting and detailed examples in the paper by Messrs. Kondo and Iwai, which reveal a number of phenomena in non-linear sampled-data systems.

B. KONDO, *in reply*

Thank you for your valuable discussion of our paper. By introducing the relation

$$y_i = x_i x_i = (x_i)$$

where  $(x)$  denotes the non-linear characteristic, eqn (5) in our paper is rewritten in the form of

$$Bx = -Af(x)$$

The above form is similar to one which you have shown, but the meaning of coefficient matrices  $A$  and  $B$  is quite different from yours.

# Continuous Compensation of Feedback Sampled-data Linear Control Systems

B. M. BROWN

## Summary

One of the more important applications of the theory of sampled data control systems is that of the design of a compensator for a feedback system when the characteristics of the plant are specified. This problem is relatively simple when a discrete compensator is used. It is, however, much more difficult in the case of a continuous compensator because of the impossibility of separating in the overall operator or transfer function the contributions of the plant and compensator operators. Such solutions as have been proposed either use approximate methods, or introduce extra samplers or subsidiary feedback loops into the system. In this paper a graphical method is described for determining the operator of a simple continuous compensator for a given plant operator, assuming that the order of the system and the poles of the overall operator are specified. It is assumed, moreover, that the compensator must be physically realizable with a plant whose load has finite inertia.

## Sommaire

Une des applications les plus importantes de la théorie des systèmes à données échantillonnées consiste à calculer un compensateur pour un système à contre réaction, les caractéristiques de l'installation étant données. Ce problème est relativement simple quand on fait usage d'un compensateur discret, mais il est beaucoup plus difficile avec un compensateur continu, parce qu'il est impossible de séparer les contributions de l'opérateur de l'installation et de celui du compensateur dans l'opérateur global. Les solutions existantes se basent sur les méthodes approximatives, ou bien elles introduisent des échantillonneurs supplémentaires ou boucles de réaction subsidiaires dans le système. Cet ouvrage décrit une méthode graphique pour déterminer l'opérateur d'un compensateur simple continu pour un opérateur d'installation donné en assumant que l'ordre du système et les pôles de l'opérateur global sont spécifiés. On assume en outre, que le compensateur doit être physiquement réalisable avec une installation dont la charge possède une inertie finie.

## Zusammenfassung

Eine der Hauptanwendungen der Theorie der Abtastregelsysteme ist die Berechnung eines Kompensationsgliedes für ein Regelsystem, wenn Eigenschaften der Anlage vorgegeben sind. Mit einem diskreten Kompensationsglied ist das Problem verhältnismäßig einfach; die Verwendung eines kontinuierlichen Kompensationsgliedes ist jedoch viel schwieriger, da sich die Gesamtübertragungsfunktion nicht in diejenige der Regelstrecke und des Kompensationsgliedes aufspalten läßt. Die bisherigen Lösungen beruhen entweder auf Näherungsverfahren oder auf der zusätzlichen Einführung von Tastern oder Rückführungen. Die Arbeit beschreibt eine graphische Methode, mit der sich die Übertragungsfunktion eines einfachen kontinuierlichen Kompensationsgliedes bei einer vorgegebenen Übertragungsfunktion der Regelstrecke für den Fall bestimmen läßt, daß die Ordnung des Systems und die Pole der Gesamtübertragungsfunktion bekannt sind. Ferner wird angenommen, daß das Kompensationsglied für eine Regelstrecke, deren Störgrößen Verzögerungen (Trägheit) besitzen, physikalisch realisierbar ist.

## Introduction

The problems of the design of compensators for linear feedback sampled-data control systems have received much attention during the last 10 years<sup>1-5</sup>. In general it has been found that when discrete compensation is specified, relatively simple mathematical processes are available to determine suitable forms and parameter values for the compensating elements. The problem is much more complicated however when continuous compensation is used. This is due basically to the fact that the equivalent difference operator or pulse transfer function of two continuous linear elements in cascade without an intervening sampler bears no simple relation to the operators or transfer functions of the separate elements. For this reason it is not easy to determine a compensator which, operating on a given plant, will provide an output with specified characteristics. The methods that have been proposed for achieving this have either been approximate, or have involved trial and error, or have entailed the introduction of extra samplers or feedback loops into the system.

This paper is perhaps not the place for a detailed discussion of the relative merits of discrete and continuous compensation. Certain of these will depend on practical rather than theoretical considerations and will vary with different applications. It can, however, be stated that with discrete compensation the signal directly controlling the plant can be changed only at the sampling instants, whereas with continuous compensation it is possible for this signal to be varied continuously. In this way it is possible at the design stage to effect greater control over the response of the system between the sampling instants. In particular, by suitable choice of the order of the system it is possible to reduce or eliminate the steady-state ripple in the output for certain specified inputs.

One effective method for designing a discrete compensator is to specify the order and characteristic equation of the system and to deduce the poles and zeros of the compensator operator<sup>3</sup>. The subject of this paper is the analogous problem with a continuous compensator. Without making any approximations it is shown how and under what conditions a physically realizable compensator of this type can be found. A special case of a problem of this type has, in fact, been discussed by Linvill and Sittler<sup>8</sup>.

## Operational Notation

Operational notation and methods are used in this paper, the basic symbols being defined as follows. As usual, the operator  $D$  applied to a function of continuous time  $t$  denotes differentiation with respect to  $t$ . When dealing with functions which are defined only for discrete equally spaced values of  $t$ , that is with sequences  $u_r$  where  $r$  takes integral values only, two basic operators are used. These are the shift and backward difference operators  $E$  and  $\nabla$ , given by the following relations:

$$Eu_r = u_{r+1}; \quad \nabla u_r = u_r - u_{r-1}$$

$$\text{so that} \quad \nabla = 1 - E^{-1} \quad (1)$$

In order to transform a function  $u(t)$  into a sequence  $u_r$ , where  $u_r = u(rT)$ , the sampling operator  $S$  is used. Thus

$$Su(t) = u_r = u(rT) \quad (2)$$

The constant  $T$  is the sampling period and in this paper is taken to have the value unity. Conversely, the first operation in the conversion of a sequence  $u_r$  into a function of  $t$  is that of pulsing, denoted by  $P$ , where

$$Pu_r = \sum_{r=-\infty}^{\infty} u_r \delta(t - rT) \quad (3)$$

$\delta(t)$  being the unit impulse function. By suitable combinations of these operators it is possible, in general, to represent unambiguously the relations between the various signals in any linear system which operates in part continuously and in part intermittently and which, except for the sampling that occurs, is time-invariant.

It is convenient to introduce two other operators. The first,  $C$  denotes the process of clamping, that is, a combination of pulsing and a zero-order hold. Thus

$$C = \nabla D^{-1} P \quad (4)$$

The second is the attenuated difference operator  $\nabla_\lambda$  defined by the relation

$$\nabla_\lambda = 1 - e^{-\lambda T} E^{-1}$$

Then if  $T = 1$ ,

$$\nabla_\lambda = e^{-\lambda} (\nabla + I) \quad (5)$$

where

$$I = e^\lambda - 1 \quad (6)$$

If a sequence  $u_r$  is pulsed and the resulting pulse train is passed through an element with operator  $F(D)$ , the output of which is sampled, the resulting sequence  $x_r$  is given by

$$x_r = SF(D) Pu_r = F_p(E) u_r$$

where

$$F_p(E) = SF(D) P \quad (7)$$

Alternatively, if the original sequence  $u_r$  is clamped, then  $x_r = F_c(E) u_r$ , where

$$F_c(E) = SF(D) C \quad (8)$$

Then

$$\frac{F_c(E)}{\nabla} = S \frac{F(D)}{D} P$$

In particular it can be shown that if  $F(D) = 1/(D + \lambda)$ , then  $F_p(E) = 1/\nabla_\lambda$ ; also that if  $F_c(E) = 1/(\nabla + I)$  a possible form of  $F(D)$  is given by

$$F(D) = e^{-\lambda} \left[ \frac{\lambda}{I(D + \lambda)} + 1 \right] \quad (9)$$

A fuller account of the use of operators in linear system analysis has been given elsewhere<sup>6, 7</sup>. It may be mentioned that if, in this paper,  $D$  is replaced by  $p$  or  $s$  and  $E$  by  $z$ , the theory is given analogously in terms of the Laplace and  $z$  transformations. In particular,  $F_p(z)$  is the pulse transfer function of an element whose transfer function is  $F(p)$ .

## General Theory

The method which will be used for synthesizing the sampling feedback control system to be considered may be appreciated more clearly if it is compared step by step with that for an equivalent continuously operating system. Consider therefore the system of *Figure 1*, with input  $u(t)$ , output  $x(t)$ , error  $c(t)$  and open-loop operator  $A(D)$ . The latter will be the product of the plant and compensator operators. It will be of the form  $P_1(D)/D^s R_1(D)$ , where  $P_1(D)$  and  $R_1(D)$  are polynomials in  $D$ . Moreover, if the load on the plant has appreciable inertia the degree of the denominator of  $A(D)$  will exceed that of the numerator by at least two. The error is given by

$$c(t) = \frac{1}{1 + A(D)} u(t) = \frac{D^s R_1(D)}{Q_1(D)} u(t) \quad (10)$$

where

$$Q_1(D) = D^s R_1(D) + P_1(D) \quad (11)$$

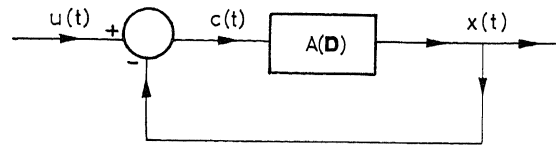


Figure 1

The degree of  $Q_1(D)$  will be the same as that of  $D^s R_1(D)$ . The stability and transient modes of the system depend on the zeros of the polynomial  $Q_1(D)$ . The significance of the constant  $s$  (called the order of the system) lies in the fact, immediately apparent from eqn (10), that the steady-state error is constant if  $D^s u(t)$  is constant and is zero if any lower order derivative of  $u(t)$  is constant.

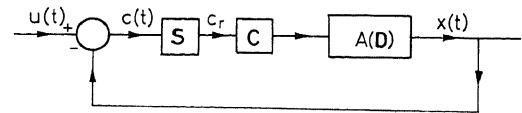


Figure 2

Consider now the system shown in *Figure 2*, in which the error is sampled and clamped before being applied to the compensator and plant. It can be shown that the error sequence is given in terms of the sampled input by the relation

$$c_r = \frac{1}{1 + A_c(E)} u_r = \frac{\nabla^s R(\nabla)}{Q(\nabla)} u_r \quad (12)$$

where, as with the continuously controlled system,

$$A_c(E) = P(\nabla)/\nabla^s R(\nabla)$$

and

$$Q(\nabla) = \nabla^s R(\nabla) + P(\nabla)$$

Then for each factor  $D + \lambda$  in  $R_1(D)$  there will be a factor  $\nabla_\lambda$  in  $R(\nabla)$ . As above, the nature of the error will depend on the order  $s$  of the system and the steady state and transient response will be determined by the zeros of the characteristic polynomial  $Q(\nabla)$ .

The problem to be considered first is as follows. Assume that these zeros, and therefore the polynomial  $Q(\nabla)$  are given.



Assume also that the poles of  $A(D)$  are known. An explicit expression for  $A(D)$  is determined and conditions deduced that it will be physically realizable for the case when the output has finite inertia. More precisely, let the poles of  $A(D)$  be  $-\lambda_1, -\lambda_2, \dots, -\lambda_k$ , and let

$$Q(\nabla) = a_0 + a_1 \nabla + \dots + a_k \nabla^k = b_0 + b_1 E^{-1} + \dots + b_k E^{-k} \quad (13)$$

Then

$$A_c(E) = \frac{Q(\nabla)}{\nabla_{\lambda_1} \nabla_{\lambda_2} \dots \nabla_{\lambda_k}} - 1$$

$$= (\exp \sum_i \lambda_i) \frac{Q(\nabla)}{\prod_i (\nabla + l_i)} - 1$$

using eqn (5).

In order to find the operator  $A(D)$  corresponding to  $A_c(E)$  and to express the restrictions which must be imposed on this operator, use is made of determinants of the type

$$\begin{vmatrix} 1 & l_1 & l_1^2 & \dots & l_1^{k-2} & F(l_1) \\ 1 & l_2 & l_2^2 & \dots & l_2^{k-2} & F(l_2) \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 1 & l_k & l_k^2 & \dots & l_k^{k-2} & F(l_k) \end{vmatrix}$$

This is denoted by  $|k, F(l)|$ . It is, in fact, a special type of alternant, that is, a determinant  $|a_{ij}|$  in which  $a_{ij} = f_j(l_i)$  where  $f_1(l), \dots, f_k(l)$  are different functions of  $l$ .

Then it is well known that  $|k, l^{k-1}|$  is the product of the differences of pairs of the numbers  $l_1, \dots, l_k$ . If  $\exp \sum_i \lambda_i$  is denoted by  $A$  it can easily be verified that the partial fraction expansion of  $A_c(E)$  can be expressed in the form

$$A_c(E) = A \left[ \frac{(-)^{k-1} |k, Q(-l)/(l+1)|}{|k, l^{k-1}|} + a_k \right] - 1 \quad (14)$$

so that, using eqn (9),

$$A(D) = A \left[ \frac{(-)^{k-1} |k, e^{-\lambda} Q(-l) \left[ \frac{\lambda}{l(D+\lambda)} + 1 \right]|}{|k, l^{k-1}|} + a_k \right] - 1 \quad (15)$$

Now if the degree of the denominator exceeds that of the numerator by at least two, the constant terms in the partial fraction expansions of  $A(D)$  and of  $DA(D)$  must both be zero. Since  $e^{-\lambda} = 1/(l+1)$  and

$$|k, F_1(l) + F_2(l)| = |k, F_1(l)| + |k, F_2(l)|$$

the first of these terms is

$$A \left[ \frac{(-)^{k-1} |k, Q(-l)/(l+1)|}{|k, l^{k-1}|} + a_k \right] - 1$$

This is found after some reduction to be  $b_0 - 1$ , so that  $b_0$  must equal 1. A direct consequence of this is that when  $A_c(E)$  is expressed as a fraction in terms of  $E^{-1}$ , its numerator contains no constant term. This result is otherwise evident since, with the type of system prescribed, the output can contain no steps. It follows from this that the value of the output at any sampling instant depends only on the value of the input at earlier sampling instants and not on its value at the same instant.

If  $b_0 = 1$ , eqn (15) reduces to

$$A(D) = \frac{(-)^{k-1} A |k, f(l)/(D+\lambda)|}{|k, l^{k-1}|} \quad (16)$$

where

$$f(l) = \lambda e^{-\lambda} Q(-l)/l = \log(l+1) Q(-l)/l(l+1)$$

It is evident that the constant term in  $DA(D)$  vanishes if

$$|k, f(l)| = 0 \quad (17)$$

This condition imposes a constraint on the possible values of  $l_i$  and therefore of  $\lambda_i$ . This is now transformed into a form in which it can be interpreted and applied by a graphical method.

If  $l_k = 0$ , the last row of  $|k, f(l)|$  becomes

$$1 \ 0 \ 0 \dots 0 \ f(0)$$

Subtracting this row from the other rows gives

$$|k, f(l)| = (-)^{k-1} l_1 l_2 \dots l_{k-1} |k-1, \{f(l) - f(0)\}/l|$$

so that eqn (17) reduces to  $|k-1, f_1(l)| = 0$  where

$$f_1(l) = [f(l) - f(0)]/l$$

More generally, if  $l_k = l_{k-1} = \dots = l_{k-s+1} = 0$ , it is found that eqn (17) is equivalent to

$$|k-s, f_s(l)| = 0 \quad (18)$$

where

$$f_s(l) = \frac{1}{l^s} \left[ f(l) - f(0) - lf'(0) - \dots - \frac{l^{s-1}}{(s-1)!} f^{(s-1)}(0) \right] \quad (19)$$

It is now shown how this equation can be solved graphically. In the simple case in which  $k-s=2$ , eqn (18) reduces to

$$f_s(l_1) = f_s(l_2) \quad (20)$$

Corresponding values of  $l_1$  and  $l_2$  can clearly be read off from a graph of  $f_s(l)$ , as shown in Figure 3.

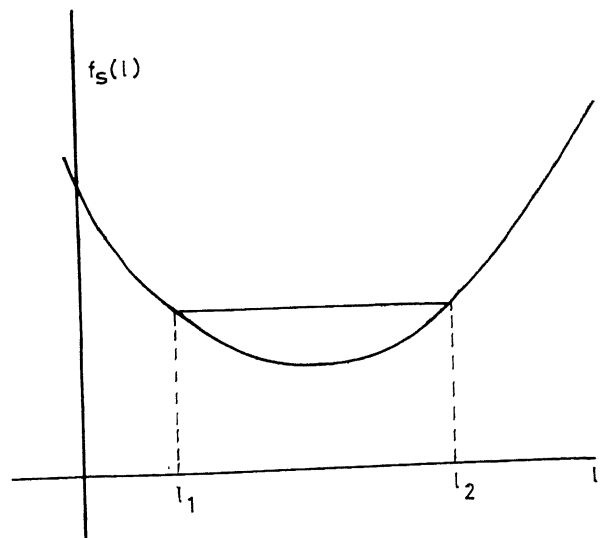


Figure 3

We consider now the case where  $k - s = 3$ , when eqn (17) takes the form

$$\begin{vmatrix} 1 & l_1 & f_s(l_1) \\ 1 & l_2 & f_s(l_2) \\ 1 & l_3 & f_s(l_3) \end{vmatrix} = 0 \quad (21)$$

This shows that the points on the graph of  $f_s(l)$  for which  $l$  takes the values  $l_1$ ,  $l_2$  and  $l_3$  must be collinear, as in Figure 4. It is evident that if two of these values are given, the third can be determined immediately.

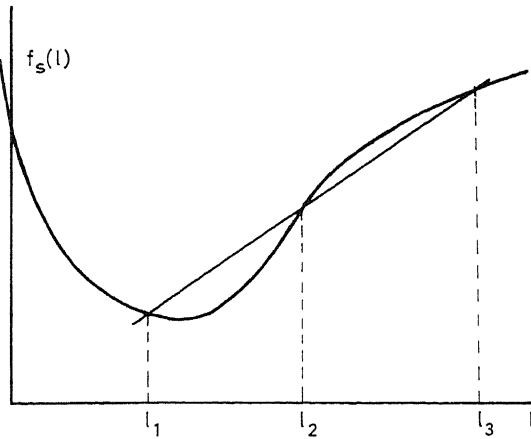


Figure 4

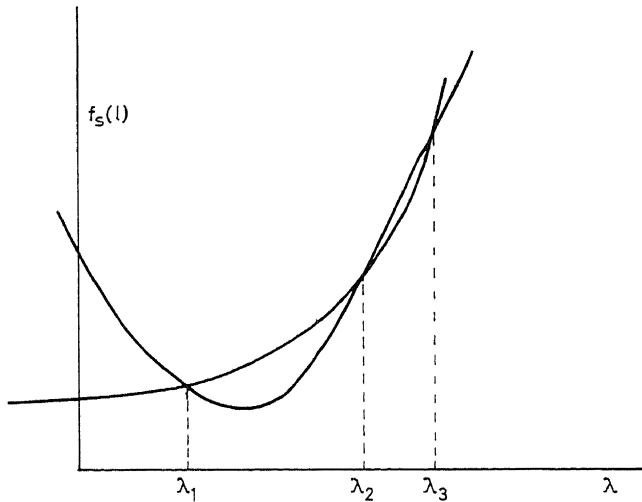


Figure 5

In practice the range of values of  $l$  that may have to be considered is such that it is preferable to plot  $f_s(l)$  against  $\lambda$  rather than  $l$ . When  $k - s = 2$  corresponding values  $\lambda_1$  and  $\lambda_2$  still give equal ordinates on the graph, but if  $k - s = 3$  corresponding points on the curve now lie on an exponential curve of the type  $y - y_0 = e^{\lambda - \lambda_0}$ , as shown in Figure 5. However, the shape of this curve is independent of  $y_0$  and  $\lambda_0$  so that  $\lambda_3$  can be found immediately from  $\lambda_1$  and  $\lambda_2$  by means of a template cut in this shape.

The cases in which  $k - s$  is equal to two or three should cover the majority of practical applications, but the graphical method used can be extended to higher values without undue

complication. Thus if  $k - s = 4$ , points on the graph of  $f_s(l)$  for related values of  $l$  lie on a parabola or on a curve with equation  $y = a e^{2\lambda} + b e^{\lambda} + c$ , depending on the choice of abscissa. In either case, if three values are given the parabola or exponential curve can be plotted and the fourth value determined by its intersection with the original graph.

### Applications

The distribution and indeed the existence of sets of related real values of  $l$  or  $\lambda$  depend on the shape of the graph of  $f_s(l)$ . This is now examined in one or two special cases.

It is found from eqns (13) and (19) that

$$f_s(l) = \frac{\log(l+1) Q(-l)}{l^{s+1}(l+1)} - \frac{a_0}{l^s} + \frac{a_1 + \frac{3}{2}a_0}{l^{s-1}} - \frac{a_2 + \frac{3}{2}a_1 + \frac{11}{6}a_0}{l^{s-2}} + \dots \quad (22)$$

the last term in this series being a multiple of  $1/l$ . Let the degree of  $Q(\nabla)$  be  $h$ . It has been assumed so far that  $h = k$ , where  $k$  is the degree of the denominator of  $A(D)$ . This is not necessarily the case. In fact, although  $h$  can never exceed  $k$ , there is no physical reason why  $h$  should not be less than  $k$ . We shall consider the case of staleness weighting, in which

$$Q(\nabla) = (1 - \alpha E^{-1})^h = (1 - \alpha)^h (1 + \beta \nabla)^h \quad (23)$$

where  $\beta = \alpha/(1 - \alpha)$  and  $\alpha$ , the staleness weighting constant, is a real number, which for stability must lie between  $\pm 1$ . The important special case in which the system has a finite settling time is given by  $\alpha = 0$ .

The potentially most useful case for detailed discussion is probably that in which  $s = 2$  and  $k \geq 4$ , that is, a system in which  $A(D)$  has a double pole at the origin and two or more non-zero poles. Now it is found in this case that if  $h < 4$ ,  $f_s(l)$  is monotonic, so that eqns (20) and (21) have no real solutions. We shall therefore take  $Q(l) = (1 - \alpha)^4 (1 + \beta l)^4$ , in which case eqn (22) becomes

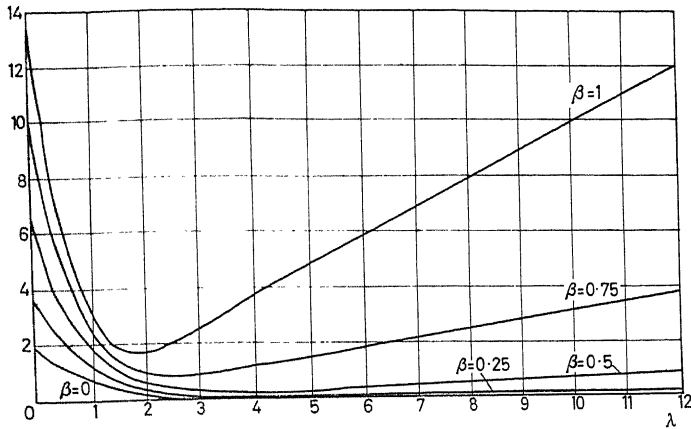
$$\frac{f_2(l)}{(1 - \alpha)^4} = \frac{\lambda(1 - \beta l)^4}{l^3(l+1)} - \frac{1}{l^2} + \frac{\frac{3}{2} + 4\beta}{l} \quad (24)$$

This function was tabulated for five different values of  $\beta$ , using an electronic computer (see Table 1 and Figure 6).

It is apparent that when  $\beta = 0$ ,  $f_2(l)$  is again monotonic, from which can be inferred the interesting result that it is not possible to achieve a finite settling time with a system of this type. With a non-zero value of  $\beta$ , however, eqns (20) and (21) become soluble.

As an example, if  $k = 4$ ,  $\alpha = \frac{1}{3}$  (or  $\beta = \frac{1}{2}$ ) and  $\lambda_1 = 3$ , it is found from Table 1 that  $\lambda_2 = 4.542$ . The subsequent calculation of  $A(D)$  is simplified somewhat if the following formulae, derived without difficulty from eqns (16) and (18) in the case when  $s = 2$ , are used.

$$A(D) = \frac{(-)^k A \left[ k-2, \frac{\lambda f(l)}{l^2 D(D+\lambda)} \right]}{|k-2, l^{k-3}|} + \frac{f(0)}{D^2} \prod_i \left[ 1 + \frac{1}{l_i} \right] \quad (25)$$


 Figure 6. Curves of  $f_2(l)/(1 - \alpha)^4$  when  $Q(\nabla) = (1 - \alpha E^{-1})^4$ 

$$\frac{\lambda f(l)}{l^2} = \lambda \left[ f_2(l) - \frac{a_1 + \frac{3}{2}a_0}{l} + \frac{a_0}{l^2} \right] \quad (26)$$

It is found that

$$A(D) = \frac{4.37D^2 + 11.42D + 2.86}{D^2(D+3)(D+4.54)} \quad (27)$$

Thus if the plant operator is  $1/D(D+3)$ , in order to satisfy the specified conditions the compensator operator would have to be

$$\frac{4.37D^2 + 11.42D + 2.86}{D(D+4.54)}$$

Alternatively, if the plant is a pure double integrator, with operator  $1/D^2$ , a suitable compensator operator could be

$$\frac{4.37D^2 + 11.42D + 2.86}{(D+3)(D+4.54)}$$

To take a more complicated example, with  $k = 5$ ,  $s = 2$  and  $\alpha = \frac{1}{2}$  (or  $\beta = 1$ ) let  $\lambda_1 = 4$  and  $\lambda_2 = 3$ . Then it is found from the graph, by using an exponential template as in Figure 5, that  $\lambda_3 = 1.3$ . If required, a more accurate value could be obtained by inverse interpolation. Substitution in eqn (25) and simplification gives

$$A(D) = \frac{1.01D^3 + 7.60D^2 + 6.92D + 1.44}{D^2(D+4)(D+3)(D+1.30)} \quad (28)$$

Thus for a plant whose operator is

$$\frac{D+2}{D(D+4)(D+3)}$$

the compensator operator to satisfy the given specification would be

$$\frac{1.01D^3 + 7.60D^2 + 6.92D + 1.44}{D(D+2)(D+1.30)}$$

### Conclusion

The theory given in this paper has been shown to provide a practicable method for solving the problem under discussion in a wide range of cases. Basically, all that is required is a table of the function  $f_s(l)$ . In fact, the time taken to prepare the programme and to produce Table 1 was less than an hour

 Table 1. Values of  $f_2(l)/(1 - \alpha)^4$  when  $Q(\nabla) = (1 - \alpha E^{-1})^4$ 

$\lambda$	$l$	$\beta = 0$ $\alpha = 0$	$\beta = 0.25$ $\alpha = 0.2$	$\beta = 0.5$ $\alpha = 0.33$	$\beta = 0.75$ $\alpha = 0.43$	$\beta = 1$ $\alpha = 0.5$
0.0	0.00000	1.8333	3.7083	6.3333	9.7083	13.8333
0.2	0.22140	1.4624	2.9059	4.8446	7.2223	9.9860
0.4	0.49182	1.1695	2.2825	3.7110	5.3731	7.1990
0.6	0.82211	0.9376	1.7974	2.8490	4.0069	5.2110
0.8	1.2255	0.7534	1.4193	2.1944	3.0060	3.8225
1.0	1.7182	0.6067	1.1239	1.6982	2.2807	2.8814
1.2	2.3201	0.4896	0.8926	1.3227	1.7624	2.2726
1.4	3.0552	0.3959	0.7111	1.0393	1.3994	1.9090
1.6	3.9530	0.3206	0.5684	0.8261	1.1523	1.7250
1.8	5.0496	0.2601	0.4558	0.6663	0.9913	1.6714
2.0	6.3890	0.2113	0.3669	0.5473	0.8943	1.7116
2.2	8.0250	0.1718	0.2964	0.4594	0.8444	1.8185
2.4	10.023	0.1399	0.2405	0.3952	0.8291	1.9720
2.6	12.463	0.1140	0.1961	0.3491	0.8389	2.1573
2.8	15.444	0.0929	0.1607	0.3168	0.8670	2.3639
3.0	19.085	0.0758	0.1325	0.2950	0.9081	2.5839
3.2	23.532	0.0619	0.1101	0.2814	0.9585	2.8120
3.4	28.964	0.0506	0.0922	0.2739	1.0153	3.0443
3.6	35.598	0.0413	0.0779	0.2711	1.0766	3.2781
3.8	43.701	0.0338	0.0665	0.2720	1.1408	3.5117
4.0	53.598	0.0276	0.0575	0.2757	1.2069	3.7440
4.2	65.686	0.0226	0.0503	0.2815	1.2741	3.9742
4.4	80.450	0.0184	0.0447	0.2889	1.3419	4.2021
4.6	98.484	0.0151	0.0403	0.2976	1.4099	4.4273
4.8	120.51	0.0123	0.0369	0.3072	1.4779	4.6500
5.0	147.41	0.0101	0.0342	0.3175	1.5457	4.8701
5.2	180.27	0.0082	0.0323	0.3284	1.6132	5.0879
5.4	220.40	0.0067	0.0308	0.3398	1.6804	5.3036
5.6	269.42	0.0055	0.0297	0.3514	1.7473	5.5173
5.8	329.29	0.0045	0.0290	0.3633	1.8138	5.7292
6.0	402.42	0.0037	0.0286	0.3753	1.8799	5.9395
6.2	491.74	0.0030	0.0284	0.3975	1.9457	6.1484
6.4	600.84	0.0024	0.0284	0.3998	2.0112	6.3560
6.6	734.09	0.0020	0.0285	0.4122	2.0764	6.5626
6.8	896.84	0.0016	0.0288	0.4246	2.1414	6.7683
7.0	1095.6	0.0013	0.0292	0.4371	2.2061	6.9731
7.2	1338.4	0.0011	0.0296	0.4495	2.2707	7.1772
7.4	1634.9	0.0009	0.0301	0.4621	2.3351	7.3807
7.6	1997.1	0.0007	0.0306	0.4746	2.3993	7.5837
7.8	2439.6	0.0006	0.0312	0.4871	2.4634	7.7862
8.0	2979.9	0.0005	0.0319	0.4996	2.5273	7.9884
8.2	3639.9	0.0004	0.0325	0.5121	2.5912	8.1902
8.4	4446.0	0.0003	0.0332	0.5247	2.6550	8.3917
8.6	5430.6	0.0002	0.0339	0.5372	2.7187	8.5930
8.8	6633.2	0.0002	0.0346	0.5497	2.7823	8.7941
9.0	8102.0	0.0001	0.0353	0.5623	2.8459	8.9951
9.2	9896.1	0.0001	0.0361	0.5748	2.9095	9.1959
9.4	12087	0.0001	0.0368	0.5873	2.9730	9.3965
9.6	14763	0.0001	0.0376	0.5998	3.0365	9.5971
9.8	18032		0.0383	0.6123	3.0999	9.7975
10.0	22025		0.0391	0.6249	3.1633	9.9979
10.2	26902		0.0399	0.6374	3.2267	10.1983
10.4	32858		0.0406	0.6499	3.2901	10.3985
10.6	40133		0.0414	0.6624	3.3534	10.5988
10.8	49019		0.0422	0.6749	3.4168	10.7990
11.0	59873		0.0429	0.6874	3.4801	10.9991
11.2	73129		0.0437	0.6999	3.5435	11.1993
11.4	89320		0.0445	0.7124	3.6068	11.3994
11.6	109096		0.0453	0.7249	3.6701	11.5995
11.8	133251		0.0461	0.7374	3.7334	11.7995
12.0	162753		0.0468	0.7499	3.7967	11.9996

and similar tables using different values of the parameters  $s$  and  $h$  could be obtained equally easily. Alternatively, if no electronic computer were available the following reduction formulae could be used:

$$lf(s+1, h, l) = f(s, h, l) - f^{(s)}(l)/s! \quad (29)$$

$$f(s, h+1, l) = -\beta lf(s-1, h, l) + f(s, h, l) \quad (30)$$

where  $f(s, h, l) = f_s(l)$  and  $Q(\nabla) = (1 - \alpha E^{-1})^h$ . Another possibility, which might well give a better response, is to choose a different form for  $Q(\nabla)$ , such as  $1 \pm \alpha^h E^{-h}$ .

The method can be extended to the case where the degree of the denominator exceeds that of the numerator by three. Possible values of  $\lambda_i$  must now satisfy the two conditions:

$$|k - s, f_s(l)| = 0, \quad |k - s, F_s(l)| = 0$$

where  $F_s(l)$  is obtained from  $\lambda f(l)$  in the same way as  $f_s(l)$  is obtained from  $f(l)$ .

A particular advantage of the graphical method is that it is possible to see from the general shape of the graph of  $f_s(l)$  whether the specification can be fulfilled at all with a compensator of a given type. This general shape can often be inferred without actual computation of  $f_s(l)$  by examining its asymptotic behaviour.

Consideration has been confined so far to systems in which all the poles of  $A(D)$  are real. The theory still holds, of course, when some of these poles are complex, and it seems possible

that a graphical method could be found for finding complex values satisfying eqns (20) and (21), if there should be a practical requirement for this.

The author gratefully acknowledges the help given by Professor J. R. Lakey and Mr. J. Flower of the Department of Nuclear Science and Technology of the Royal Naval College, in the preparation of Table 1.

## References

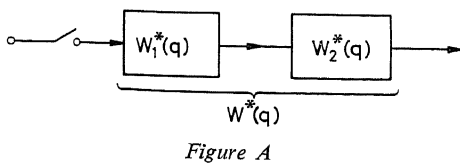
- JURY, E. I. *Proc. 1st I.F.A.C. Congr.* p. 262. 1961. London; Butterworths
- JURY, E. I. *Sampled-Data Control Systems*. 1958. New York; Wiley
- RAGAZZINI, J. R. and FRANKLIN, G. F. *Sampled-Data Control Systems*. 1958. New York; McGraw-Hill
- TOU, J. T. *Digital and Sampled-Data Control Systems*. 1959. New York; McGraw-Hill
- TSYPKIN, YA. Z. *Theory of Pulse Systems*. 1959. Moscow; Fizmatgiz
- BROWN, B. M. *Proc. 1st I.F.A.C. Congr.* p. 270. 1961. London; Butterworths
- BROWN, B. M. *The Mathematical Theory of Linear Systems*. 1961. London; Chapman and Hall
- LINVILL, W. K. and SITTNER, R. W. Extension of conventional techniques to the design of sampled-data systems. *Inst. Radio Engrs Conv. Rec.* p. 95. 1953. New York

## DISCUSSION

YA. Z. TSYPKIN, *Institute of Automation and Telemechanics, Moscow I-53, Kalachevskaya 15-A, U.S.S.R.*

Professor Brown's interesting paper examines the problem of synthesis of sampled-data systems with continuous compensation. Very few studies have been made of this problem.

The difficulties involved in solving it stem from the fact that for the continuous part (Figure A)



$$W^*(q) \neq W^*(q) \cdot W^*(q)$$

( $q = pT$ ;  $T$  is the repetition period)

But, as shown<sup>1</sup>,

$$W^*(q) = \int_0^1 W_1^*(q, -\lambda) \cdot W_2^*(q, \lambda) d\lambda \quad (1)$$

It seems that this correlation (1) will make it possible to solve the problem under consideration by the same method as is used in the synthesis of pulse-correction systems. I should like to draw Professor Brown's attention to this possibility.

## Reference

- TSYPKIN, YA. Z. *Theory of Linear Sampled-data Systems*. Moscow, 1963

R. E. KALMAN, *Research Institute for Advanced Studies, 7212 Bellona Ave., Baltimore 12, MD., U.S.A.*

The problem discussed in this paper is an interesting one in the framework of early sampled-data theory. But today transform methods are seldom used for such problems because the state-variable approach is much simpler. And today one does not design by trial and error because optimal design is easier to explain and to compute. In the light of modern control theory, the solution of the problem becomes trivial.

This solution, as is well known<sup>1</sup>, consists of two parts: (1) Determine the optimal linear control law  $K(X)$  as a function of the state variables  $X$  (Reference 2, below); (2) Determine the optimal estimates  $\hat{X}$  of the state variables  $X$  by a linear Wiener-Kalman filter<sup>2</sup>.

These two steps may be carried out completely independently of each other<sup>1</sup>, and therefore either or both may be solved in the continuous time or sampled-data (discrete-time) sense. In any case, the optimal control law is  $K(\hat{X})$ <sup>1</sup>.

## References

- JOSEPH, P. and TOU, J. *Trans. Amer. Inst. elect. Engrs.*, Pt. II (1961)
- KALMAN, R. E. and KOECKE, R. W. *Trans. Amer. Soc. mech. Engrs.* (1958)
- KALMAN, R. E. and BUCY, R. S. *Trans. Amer. Soc. mech. Engrs.; J. Bas. Engr.* (1961)

G. R. CARROLL, *University of Manchester (Faculty of Technology), Allen Hall, Wilmslow Road, Manchester 14*

Following Dr. R. E. Kalman's remarks that there is a need for the optimal control theory approach to this problem of the design of a continuous compensator for a feedback sampled-data linear control

system, as opposed to the two methods suggested here by Brown and Tsypkin, I would like to state that I have already successfully carried out such a design synthesis procedure under Dr. J. B. Knowles at the College of Science and Technology, Manchester, England.

The object was to obtain the optimum continuous passive network of any discrete compensator. The criterion of synthesis used was that

of minimum integral error squared. Satisfactory closed forms of the error function were obtained, and were then applied to a digital computer for the optimum constrained solutions of all of the independent variables of the continuous passive network compensator.

Full details of this optimal design synthesis of a continuous compensator for a sampled-data system are soon to be published.

# Fundamentals of the Theory of Non-linear Pulse Control Systems

Ya. Z. TSYPKIN

## Summary

Non-linear pulse control systems which consist of a non-linear element and a linear pulse part (LP) are considered. Using the ideas of V. M. Popov, the sufficient criterion of absolute stability is established from the frequency characteristic of the linear pulse part.

A class of non-linear pulse control systems is defined for which this criterion of stability is also necessary. The applicability of the criterion of absolute stability is shown for non-linear pulse control systems in which the characteristic of the non-linear element is an explicit time function.

It is established in which cases the absence of periodic conditions guarantees the absolute stability of these non-linear pulse control systems. Methods are given for estimating indices of the quality of a process: the degree of stability and the overall square deviation. The stability and quality of non-linear pulse control systems can be investigated comparatively simply by concepts and representations of the theory of linear pulse systems.

## Sommaire

On considère des systèmes asservis échantillonnés nonlinéaires comprenant une partie nonlinéaire et une partie échantillonnée linéaire. A partir des idées de V. M. Popov, on établit la condition suffisante de stabilité absolue à partir des caractéristiques fréquentielles de la partie échantillonnée linéaire.

On définit une classe de systèmes échantillonnés nonlinéaires pour lesquels ce critère de stabilité est également nécessaire. On montre l'applicabilité du critère de stabilité absolue à ceux dans lesquels l'élément nonlinéaire est une fonction explicite du temps.

On établit dans quels cas l'absence de conditions périodiques garantit la stabilité absolue de ces systèmes. On donne des méthodes pour évaluer les indices de qualité d'un processus: le degré de stabilité et l'écart quadratique global. La stabilité et la qualité de ces systèmes peuvent être étudiées assez simplement en partant de la théorie des systèmes échantillonnés linéaires.

## Zusammenfassung

Nichtlineare Abtastregelsysteme, welche aus einem nichtlinearen Element und einem linearen Impulsteil (LP) bestehen, werden hier besprochen. Das ausreichende Kriterium für absolute Stabilität wird nach V. M. Popov auf Grund der Frequenzcharakteristik des linearen Impulsteiles aufgestellt.

Es wird eine Gruppe nichtlinearer Abtastsysteme definiert, für die dieses Stabilitätskriterium ebenfalls notwendig ist. Die Anwendbarkeit des Kriteriums der absoluten Stabilität wird für nichtlineare Abtastregelsysteme gezeigt, bei denen die Charakteristik des nichtlinearen Elementes eine explizite Zeitfunktion ist.

In bestimmten Fällen garantiert das Fehlen periodischer Bedingungen die absolute Stabilität derartiger Regelsysteme. Es werden Methoden für die Abschätzung der Qualitätskennzeichen eines Prozesses angegeben: der Stabilitätsgrad und die gesamte quadratische Abweichung. Stabilität und Güte solcher Systeme können verhältnismäßig einfach mit Hilfe von Begriffen und Darstellungsweisen aus der Theorie nichtlinearer Abtastsysteme untersucht werden.

## Introduction

The theory of linear pulse control systems has attained a high level of development and the main problems in the analysis and synthesis of such systems can be solved. However, with regard to non-linear pulse control systems, the theory is still in its initial stage. Up to the present time non-linear theory has been confined mostly to the investigation of periodic regimes. Yet periodic regimes are not operational conditions and the important problem still remains to ensure the stability of non-linear pulse control systems and to assess the 'quality' of stable processes. Attempts to employ the methods of investigating periodic regimes for estimating stability when the required periodic regimes are no longer present are often unjustified since the absence of a particular type of periodic regime is no guarantee that other forms of periodic or almost-periodic regimes are not present.

For solution of the stability problem it was quite natural to try to employ the ideas of Liapunov's second method which is widely used in the theory of continuous systems, in extending them to difference equations<sup>1-5</sup>.

However, such an approach involves difficulties associated with the need to transform the equations of non-linear pulse systems into their normal form, the arbitrariness of the selection of Liapunov functions and the impossibility of establishing any general properties of non-linear pulse control systems.

The approach to the problem in this paper is based on an idea which Popov<sup>6, 7</sup> used in the investigation of non-linear continuous control systems. The distinctive feature of this approach is that it is closely associated with such physical concepts as the frequency and transient responses, and it provides the widest sufficient conditions of stability which can be obtained by all the Liapunov functions of the quadratic type. This approach greatly simplifies an assessment of the quality of processes in non-linear pulse control systems. It is possible to establish when the absence of periodic regimes guarantees stability and, finally, use may be made of methods similar to those employed in the investigation of linear pulse systems.

## Statement of the Problem

A block diagram of a non-linear pulse control system is shown in Figure 1. It consists of a non-linear element in series with a linear impulse part LP which is an open linear

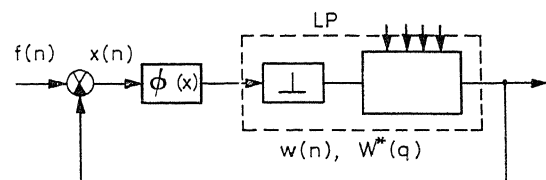


Figure 1

pulse loop. The linear impulse part incorporates a pulse element for amplitude modulation of arbitrarily shaped pulses and a continuous part.

Let us suppose that the characteristic  $\Phi(x)$  of the non-linear element satisfies the following conditions (Figure 2):

- (a)  $\Phi(0) = 0$
- (b)  $0 < \frac{\Phi(x)}{x} < k$
- (c)  $\lim_{x \rightarrow \pm \infty} \Phi(x) = \pm 0$

which correspond to the fact that this characteristic belongs to the sector  $(0, k)$ .

The main problem is to determine the stability of the systems which are to be considered for any initial deviations and to determine the quality of behaviour in stable systems. Stability of this kind which is independent of the particular shape of the characteristic of the non-linear element and which satisfies the general conditions (1), is called generally absolute stability<sup>8</sup>.

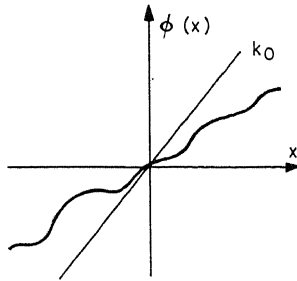


Figure 2

### Equations of Non-linear Pulse control Systems

Let one suppose that the continuous part of the linear pulse part  $LP$  receives perturbations in the form of initial conditions with  $n = 0$ . One puts  $f[n]$  for the response of this continuous part to partial conditions and applies it to the input of the non-linear pulse system (Figure 1). If the continuous part, and therefore the linear pulse part, is stable, then

$$\lim_{n \rightarrow \infty} f[n] = 0 \quad (2)$$

The equation of the pulse control system with respect to the error  $x[n]$  can take either of two forms.

(i) With respect to original lattice functions:

$$x[n] = f[n] - \sum_{m=0}^n w[n-m] \Phi(x[m]) \quad (3)$$

(ii) With respect to their transforms:

$$X^*(q) = F^*(q) - W^*(q) D \{ \Phi(x[n]) \} \quad (4)$$

Here<sup>9</sup>

$$Z^*(q) = D \{ z[n] \} = \sum_{n=0}^{\infty} e^{-qn} z[n] \quad (5)$$

is the discrete Laplace transform ( $D$  transformation);  $q = \sigma + j\bar{\omega}$  is a parameter of transformation;  $\bar{\omega} = \omega T$  is the relative frequency;  $T$  is the repetition interval;

$$W^*(q) = D \{ w[n] \} \quad (6)$$

is the transfer function of the linear pulse part;

$$w[n] = w(\bar{t})_{\bar{t}=n} \quad (7)$$

is the impulse characteristic of the linear pulse part;  $x[n]$ ,  $f[n]$  are the lattice functions, which correspond to the error and the input:  $X^*(q)$ ,  $F^*(q)$  are their transforms and, finally,  $\Phi(x)$  is the characteristic of the non-linear element.

For a stable linear pulse part

$$\lim_{n \rightarrow \infty} w[n] = 0 \quad (8)$$

This implies that the corresponding transfer function  $W^*(q)$  has no poles in the right-hand half-band  $\text{Re } q \geq 0$ ,  $-\pi < \text{Im } q \leq \pi$ .

### The Sufficient Condition of Absolute Stability

A pulse control system is absolutely stable relative to any perturbation  $f[n]$  which satisfies the condition (2) if

$$\lim_{n \rightarrow \infty} x[n] = 0 \quad (9)$$

In order to establish the fact of absolute stability, one estimates the solutions  $x[n]$  of the equation with respect to the original functions.

By analogy with the ideas of Popov<sup>6, 7</sup>, the auxiliary functions are introduced

$$\varphi_N[n] = \begin{cases} \Phi(x[n]) & 0 \leq n \leq N \\ 0 & n < 0, n > N \end{cases} \quad (10)$$

and

$$\psi_N[n] = x_N[n] - \frac{1}{k} \varphi_N[n] \quad (11)$$

where

$$x_N[n] = f[n] - \sum_{m=0}^n w[n-m] \varphi_N[m] \quad (12)$$

It is obvious that for  $0 \leq n \leq N$

$$x_N[n] \equiv x[n]$$

where  $x[n]$  is the solution of eqn (3).

Now the following expression is formed

$$\rho_N = \sum_{n=0}^{\infty} \varphi_N[n] \psi_N[n] \quad (13)$$

which, having regard to (10) and (11), is equal to

$$\rho_N = \sum_{n=0}^{\infty} \left( \Phi(x[n]) x[n] - \frac{1}{k} \Phi^2(x[n]) \right) \quad (14)$$

According to the Liapunov-Parseval equality<sup>9</sup> eqn (13) can also be represented as

$$\rho_N = \frac{1}{2\pi} \int_{-\pi}^{\pi} \Phi_N^*(-j\bar{\omega}) \Psi_N^*(j\bar{\omega}) d\bar{\omega} \quad (15)$$

where

$$\Phi_N^*(j\bar{\omega}) = D \{ \varphi_N[n] \}_{q=j\bar{\omega}} \quad (16)$$

and by virtue of (11) and (12)

$$\Psi_N(j\bar{\omega}) = D \{ \psi_N[n] \}_{q=j\bar{\omega}} = F^*(j\bar{\omega}) - \left( W^*(j\bar{\omega}) + \frac{1}{k} \right) \Phi_N^*(j\bar{\omega}) \quad (17)$$

These spectral functions exist if conditions (10) and (8) are fulfilled.

Substituting (16) and (17) into (15) and after simple transformations one gets

$$\rho_N = -\frac{1}{2\pi} \int_{-\pi}^{\pi} \left| \sqrt{\operatorname{Re} \Pi^*(j\bar{\omega})} \Phi_N^*(j\bar{\omega}) - \frac{F^*(j\bar{\omega})}{2\sqrt{\operatorname{Re} \Pi^*(j\bar{\omega})}} \right|^2 d\bar{\omega} + \frac{1}{8\pi} \int_{-\pi}^{\pi} \frac{F^*(j\bar{\omega})}{\operatorname{Re} \Pi^*(j\bar{\omega})} d\bar{\omega} \quad (18)$$

where

$$\operatorname{Re} \Pi^*(j\bar{\omega}) = \operatorname{Re} W^*(j\bar{\omega}) + \frac{1}{k} > 0 \quad (19)$$

The function

$$\Pi^*(j\bar{\omega}) = W^*(j\bar{\omega}) + \frac{1}{k}$$

which plays the main role, is called the analogue of the Popov's function.

Since the first integral in (18) is negative, by discarding it, one obtains the inequality

$$\rho_N \leq \frac{1}{8\pi} \int_{-\pi}^{\pi} \frac{|F^*(j\bar{\omega})|^2}{\operatorname{Re} \Pi^*(j\bar{\omega})} d\bar{\omega} = C \quad (20)$$

By virtue of (19) the quantity  $C$  is positive: it is independent of  $N$ .

Substituting into the left-hand side of (20) the value of  $\rho_N$  from (14) one obtains

$$\sum_{n=0}^N \Phi(x[n]) x[n] \left( 1 - \frac{\Phi(x[n])}{k x[n]} \right) \leq C \quad (21)$$

According to the condition (1a), the sum on the left-hand side of (21) is positive, moreover it is limited. The series which is formed from this sum as  $N \rightarrow \infty$ , therefore converges. Using the known theorem of the convergence of series with positive terms, one concludes that

$$\lim_{n \rightarrow \infty} \Phi(x[n]) x[n] \left( 1 - \frac{\Phi(x[n])}{k x[n]} \right) = 0$$

Hence, by virtue of the conditions (1), it follows that

$$\lim_{n \rightarrow \infty} x[n] = 0 \quad (22)$$

Thus a pulse control system which has a stable pulse linear part and a non-linear characteristic  $\Phi(x)$  and which satisfies the conditions (1), will be absolutely stable if the real part of the analogue of Popov's function is positive, i.e. if

$$\operatorname{Re} \Pi^*(j\bar{\omega}) = \operatorname{Re} W^*(j\bar{\omega}) + \frac{1}{k} > 0 \quad (23)$$

The condition of stability (23) determines the magnitude of the sector  $(0, k)$  which includes the non-linear characteristic  $\Phi(x)$  for which the pulse system is absolutely stable. This condition is sufficient.

### Frequency Criteria of Absolute Stability

To formulate the criteria of stability of a pulse control system one introduces the concept of a static gain of the non-linear element

$$S(x) = \frac{\Phi(x)}{x} \quad (24)$$

which is the slope of a straight line passing through the origin and the point of the non-linear characteristic for a specified value of  $x$ . The maximum  $S_{\max}$  and the minimum  $S_{\min}$  static gains are determined by the rays of a sector which is tangential to the characteristic (Figure 3). A non-linear pulse control system in which the non-linear element is replaced by a linear element with some fixed gain,  $k$ , is said to be a linearized pulse control system. For a linearized pulse system to be stable, by analogy with the Nyquist criterion<sup>7</sup>, it is necessary and sufficient that the frequency characteristic of the linear pulse part LP should not embrace the point  $-1/k, j0$ . It will be said that a linearized system is obviously stable if the frequency characteristic of the linear pulse part does not intersect the straight line  $-1/k$ . Then, according to the condition of stability (23), the frequency criterion of absolute stability of a non-linear pulse control system can be formulated in the following way.

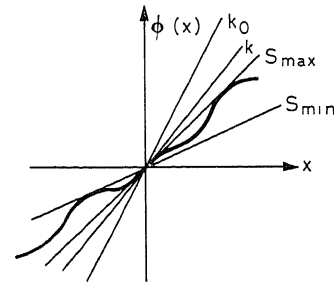


Figure 3

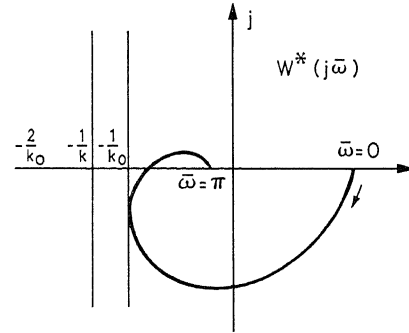


Figure 4

A non-linear pulse control system with its characteristic belonging to the sector  $(0, k)$ , will be absolutely stable if the linearized pulse system corresponding to it is obviously stable or if the frequency characteristic  $W^*(j\bar{\omega})$  of the linear pulse part does not intersect the straight line  $-1/k$  (Figure 4).

The greatest value  $k = k_0$  which determines the span of the sector in which the non-linear characteristic is located, is determined by drawing the vertical tangent to  $W^*(j\bar{\omega})$ . The difference  $k - S_{\max}$  characterizes the margin of stability.

The stability criterion of a pulse control system can also be formulated with reference to the frequency characteristic  $K^*(j\bar{\omega})$  of a closed linearized pulse control system. Selecting  $k = k_0/2$ ; then

$$K^*(j\bar{\omega}) = \frac{\frac{k_0}{2} W^*(j\bar{\omega})}{1 + \frac{k_0}{2} W^*(j\bar{\omega})} \quad (25)$$



According to the usual constructions of the frequency characteristic of a closed loop from the frequency characteristic of an open loop<sup>9</sup>, for an obviously stable linearized pulse control system if  $k = k_0/2$ , one has

$$|K^*(j\bar{\omega})| \leq 1 \quad (26)$$

Thus a non-linear pulse control system with its characteristic belonging to the sector  $(0, k_0)$  will be absolutely stable if the frequency characteristic of the closed linearized pulse control system  $K^*(\bar{\omega})$  with gain  $k_0/2$  does not exceed unity in absolute value.

One notes that the frequency criteria are also applicable in those cases when the continuous part contains delay elements or elements with distributed constants.

The frequency criteria of absolute stability can also be expressed in analytic form. The first criterion is closely related to the problem of Karateodory, whilst the second criterion is closely associated with Shur's problem in the theory of analytic functions<sup>10</sup>.

The analytic form of the criteria is considered in a special paper. One will not consider it here as, more over, the use of frequency criteria is the simplest way of elucidating various general properties of non-linear pulse control systems.

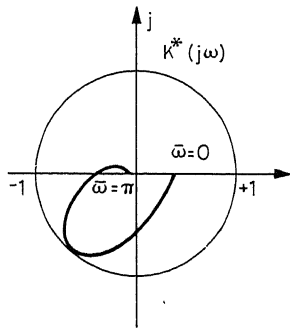


Figure 5

### Generalization of the Stability Criteria

Non-linear pulse control system which contain a stable linear pulse part have been considered above. Now suppose that the linear pulse part is neutral or unstable. This implies that its transfer function  $W^*(q)$  has poles on the imaginary axis, and in particular, at the origin or in the right-hand half-band  $\text{Re } q \geq 0$ ,  $-\pi < \text{Im } q \leq \pi$ <sup>9</sup>. Since the determined sufficient conditions must hold for any non-linear characteristic which belongs to the sector  $(0, k_0)$ , they must also hold for a linear characteristic which belongs to this sector. But for sufficiently small gains  $z$  of this linear characteristic, a closed pulse control system will behave like an open pulse control system corresponding to the linear pulse part, i.e. it will be neutral or unstable. Therefore, for instances of a neutral or unstable linear pulse part it is necessary to impose additional limitations on the minimum static gain  $S_{\min}$ . Let us elucidate these limitations. Let us introduce proportional feedback with the coefficient  $z$  across the linear pulse part (Figure 6), one supposes that the structure of the linear pulse part is such that for a finite  $z < S_{\min}$  the closed linear pulse part is stable. The frequency criteria of stability are then applicable to this non-linear pulse control system, but the role of the frequency characteristic of the

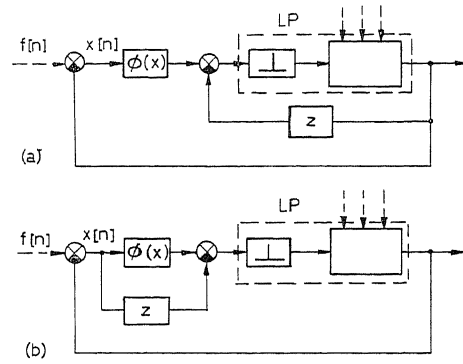


Figure 6

linear pulse part  $W^*(j\bar{\omega})$  will now be played by the frequency characteristic of the closed pulse control system, which is a new linear pulse part equal to

$$W_e^*(j\bar{\omega}) = \frac{W^*(j\bar{\omega})}{1 + r W^*(j\bar{\omega})} \quad (27)$$

But the block diagram of a non-linear pulse control system [Figure 6(a)] can easily be converted to the form of Figure 6(b) where  $f[n]$  is now the response of the closed linear pulse control system, and the non-linear characteristic is equal to

$$\Phi(x) - zx \quad (28)$$

However, since this characteristic must satisfy the conditions (1),

$$r < \frac{\Phi(x)}{x} < k_1 + z = K \quad (29)$$

i. e.

$$S_{\min} > z$$

Thus the formulation of the frequency criterion remains unchanged. Only the characteristic of the non-linear element must now belong to the sector  $(z, k+z)$ , and the frequency characteristic of the linear pulse part  $W_e^*(j\bar{\omega})$  is determined by the expression (27).

One notes that if the linear pulse part is neutral and its transfer function  $W^*(q)$  has only one pole at the origin, whilst the rest of the poles have negative real parts, then  $z$  in eqn (27) can be arbitrarily small and for this case one has

$$W_e^*(j\bar{\omega}) \approx W^*(j\bar{\omega}) \quad (30)$$

i.e. in this case there is no need to construct  $W_e^*(j\bar{\omega})$  from  $W^*(j\bar{\omega})$  on the basis of the relation (27).

If the non-linear characteristic  $\Phi(x)$  at  $x \geq x^0$  goes outside the limits of the sector  $(z, k+z)$ , which is usually the case for non-linear characteristics of the saturation type, the frequency criterion of stability guarantees stability with deviations of the error not exceeding  $x^0$ .

The frequency criteria of stability also hold for those cases when the non-linear characteristic (or gain of the linear pulse part) is a function of time  $n$ , if  $\Phi(x, n)$  for any  $n \geq n_0$  satisfies the conditions (1), i.e. if it belongs to the sector  $(0, k_0)$  or in the case of a neutral or unstable linear pulse part belongs to the sector  $(z, k_0 + z)$ .

### The Necessary and Sufficient Conditions of Absolute Stability for Some Non-linear Control Systems

Frequency criteria of absolute stability determine the sufficient conditions of absolute stability. It is obvious that in those cases when these sufficient conditions of absolute stability coincide with the necessary and sufficient condition of stability of linearized pulse control systems, they also become necessary conditions of absolute stability. Let us define the class of non-linear pulse control systems for which the conditions of absolute stability are necessary and sufficient. This problem was first posed by Aizerman<sup>11</sup>, for continuous control systems, and slightly later by Letov<sup>8</sup>. The solution of this problem is of importance since it permits reduction of the investigation of the absolute stability of non-linear pulse control systems to the well-known investigation of the stability of linear pulse control systems.

It follows directly from the formulation of the frequency criterion that this class of non-linear pulse control systems includes those for which the obvious stability of linearized pulse control systems coincides with their stability. The frequency characteristics of these latter pulse control systems  $W^*(j\bar{\omega})$  [or  $W_e^*(j\bar{\omega})$ ] must have the form shown in Figure 7(a) and (b).

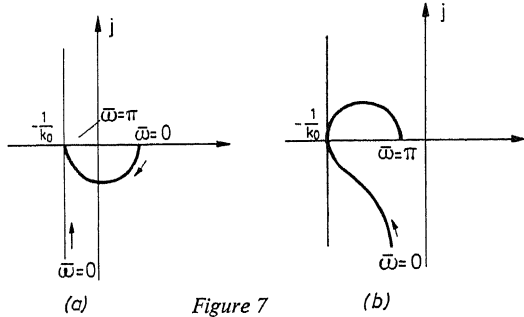


Figure 7

The frequency criterion of absolute stability determines the necessary and sufficient conditions for all non-linear pulse control systems of the first order (with amplitude- or pulse width- or time-modulation), and also for non-linear pulse control systems of any order whose frequency characteristic  $W^*(j\bar{\omega})$  has the largest real part in absolute value at the boundary frequency. It is worthwhile pointing out that for this class of system the absence of periodic conditions according to the improved method of harmonic balance<sup>12</sup>, testifies to their stability. For digital automatic systems, as shown elsewhere<sup>13</sup>, the determination of periodic regimes with a relative frequency  $\bar{\omega} = \pi$  entails drawing a straight line with a slope  $-1/W^*(j\bar{\omega})$  in the plane of the non-linear characteristic (Figure 8)\*. If the maximum real part  $W^*(j\bar{\omega})$  in absolute magnitude is attained for  $\bar{\omega} = \pi$  (which always occurs for first-order pulse control systems), the condition requiring the absence of periodic regimes with a relative frequency  $\bar{\omega} = \pi$  coincides with the condition of absolute stability.

### Estimation of the Degree of Stability

For the simplest estimate of the quality of the behaviour of a non-linear pulse control system, one will use the concept of degrees of stability which characterizes the process damping speed.

\* The author points out that in a previous paper<sup>12</sup> he has given an erroneous slope.

For this purpose, instead of the auxiliary functions (10) and (11), the following functions are introduced.

$$\varphi_N[n] = \begin{cases} \Phi(x[n]) e^{\delta n} & 0 \leq n \leq N \\ 0 & n < 0, n > N \end{cases} \quad (31)$$

and

$$\psi_N[n] = x_N[n] e^{\delta n} - \frac{1}{k} \varphi_N[n] \quad (32)$$

where  $\delta > 0$  is some constant quantity.

Multiplying both sides of (12) by  $e^{\delta n}$ , there is obtained

$$x_N[n] e^{\delta n} = f[n] e^{\delta n} - \sum_{m=0}^n w[n-m] e^{\delta(n-m)} \psi_N[m] \quad (33)$$

Remarking that according to the shift theorem<sup>9</sup>

$$D\{z[n] e^{\delta n}\}_{q=j\bar{\omega}} = Z^*(-\delta + j\bar{\omega}) \quad (34)$$

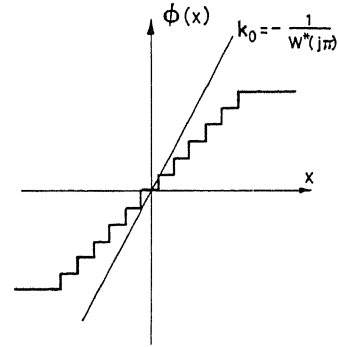


Figure 8

and following the same discussion as in the establishment of the condition of absolute stability, the conclusion is reached that

$$\lim x[n] e^{\delta n} = 0 \quad (35)$$

if the real part of the analogue of the shifted function of Popov is positive, i.e. if

$$\operatorname{Re} \Pi^*(-\delta + j\bar{\omega}) = \operatorname{Re} W^*(-\delta + j\bar{\omega}) + \frac{1}{k} > 0 \quad (36)$$

As will be seen from eqn (35), the rate of damping is determined here by the quantity  $\delta$ . The determination of the conditions for which non-linear pulse control systems have a specified degree of stability,  $\delta_0$ , thus entails the use of the frequency criterion of stability and its application to the shifted frequency characteristic

$$W^*(-\delta_0 + j\bar{\omega}) \quad (37)$$

or

$$W_e^*(-\delta + j\bar{\omega}) = \frac{W^*(-\delta_0 + j\bar{\omega})}{1 + z W^*(-\delta_0 + j\bar{\omega})} \quad (38)$$

for a fixed value  $\delta_0$  (Figure 9).

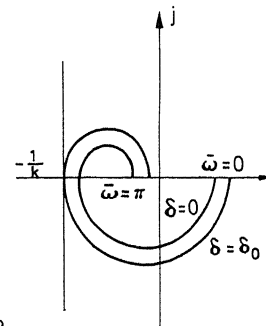


Figure 9

Since the poles of the transfer function  $W^*(-\delta + q)$  depend on  $\delta$  and with increase of  $\delta$  are shifted in the direction of the right-hand half-band, the greatest value of  $\delta = \delta_{\max}$  is attained for a value

$$z_0 \leq S_{\min} \quad (39)$$

which still ensures stability of a closed linear pulse part. Thus the increase of  $\delta$  is possible until the poles  $W^*(-\delta + q)$  are located at the origin or on the imaginary axis. With an increase of  $\delta$  the quantity  $k_0$  usually decreases, whilst  $z$  increases. Therefore, the less the difference  $S_{\max} - S_{\min}$ , the more attainable is a large degree of stability. This estimate is also applicable to non-linear pulse control systems in which the characteristic of the non-linear element depends also on time.

### The Quadratic Estimate

Another important estimate of the quality of behaviour is the overall quadratic estimate of the output of a non-linear element

$$I_2 = \sum_{n=0}^{\infty} \Phi^2(x[n]) \quad (40)$$

To determine the upper boundary of this estimate, one will avail oneself of Popov's ideas<sup>14</sup>. Consider the inequality (21) for  $N = \infty$ , representing it in the form

$$\rho_x = \sum_{n=0}^{\infty} \Phi^2(x[n]) \left( \frac{x[n]}{\Phi(x[n])} - \frac{1}{k} \right) \leq C \quad (41)$$

Since

$$\frac{x}{\Phi(x)} \geq \frac{1}{S_{\max}}$$

the inequality (41) can be strengthened and

$$\left( \frac{1}{S_{\max}} - \frac{1}{k} \right) \sum_{n=0}^{\infty} \Phi^2(x[n]) \leq C \quad (42)$$

Taking into account the notation of (40) and (20), from (42) one gets

$$I_2 \leq \frac{k S_{\max}}{k - S_{\max}} \frac{1}{8\pi} \int_{-\pi}^{\pi} \frac{|F^*(j\bar{\omega})|^2}{\operatorname{Re} \Pi^*(j\bar{\omega})} d\bar{\omega} \quad (43)$$

where

$$\operatorname{Re} \Pi^*(j\bar{\omega}) = \operatorname{Re} W^*(j\bar{\omega}) + \frac{1}{k} \geq \frac{1}{k} - \frac{1}{k_0} > 0 \quad (44)$$

Replacing  $\operatorname{Re} \Pi^*(j\bar{\omega})$  in eqn (43) by its maximum value, one finally gets

$$I_2 \leq \frac{k_0 k^2 S_{\max}}{(k - S_{\max})(k_0 - k)} \frac{1}{8\pi} \int_{-\pi}^{\pi} |F^*(j\bar{\omega})|^2 d\bar{\omega} \quad (45)$$

The right-hand side of inequality (44) contains an undetermined parameter  $k$ ; here (Figure 3)

$$S_{\max} < k < k_0 \quad (46)$$

This is so selected that the coefficient is minimum for the integral (45). It can be shown without difficulty that in this case

$$k = \frac{2k_0 S_{\max}}{k_0 + S_{\max}} \quad (47)$$

and therefore finally get

$$I_2 \leq \frac{k_0^2 S_{\max}^2}{(k_0 - S_{\max})^2} \frac{1}{2\pi} \int_{-\pi}^{\pi} |F^*(j\bar{\omega})|^2 d\bar{\omega} \quad (48)$$

But according to the Liapunov-Parseval<sup>9</sup> equality,

$$\frac{1}{2\pi} \int_{-\pi}^{\pi} |F^*(j\bar{\omega})|^2 d\bar{\omega} = \sum_{n=0}^{\infty} f^2[n] \quad (49)$$

Therefore eqn (48) can also be represented as

$$I_2 \leq \frac{k_0^2 S_{\max}^2}{(k_0 - S_{\max})^2} \sum_{n=0}^{\infty} f^2[n] \quad (50)$$

It follows from (50) that the upper boundary of the overall quadratic estimate is determined by the sum of the squares of the discrete responses of the linear impulse part to the applied inputs. If the linear impulse part receives an input  $f_1[n]$  which decreases with time, then

$$f[n] = \sum_{m=0}^n w[n-m] f_1[m]$$

and this implies that<sup>9</sup>

$$F^*(j\bar{\omega}) = D \{f[n]\}_{q=j\bar{\omega}} = \mathfrak{D} \{W^*(q) F_1^*(q)\}_{q=j\bar{\omega}}$$

The computation of the right-hand sides of (48) or (50) is carried out analytically or graphically by known rules<sup>9</sup>. The upper boundary of the overall quadratic error is less, other things being equal, the greater the margin of stability  $k_0 - S_{\max}$ . This estimate is also applicable when the characteristic of a non-linear element depends also on time.

### Conclusion

This approach to the problem makes it comparatively simple by the concepts of the linear theory of pulse systems to determine the region of absolute stability of non-linear pulse control systems and to estimate indices of the quality of processes (the degree of stability and the overall quadratic estimate). The fact that the stability and estimates of indices of process quality are independent of the actual shape of the characteristic of the non-linear element, provided only that this characteristic belongs to the specified sector, makes it possible to ensure values of estimates of the indices of quality for variation of the characteristic of the non-linear element or of the parameters of the linear pulse part which also lead to a change in the boundaries of the sector ( $z, k+z$ ). In some cases it is therefore no longer necessary to use special additional self-adjusting circuits which complicate non-linear pulse control systems.

In this connection it is extremely important to determine the structure of non-linear pulse control systems, the sensitivity<sup>15</sup> of which is low in relation to variations of the non-linear characteristic and to the parameters of the linear part. For this purpose use may be made of the results of investigations into the sensitivity of linear pulse control systems.

Generalization of the method of investigating non-linear pulse control systems to pulse control systems which contain a linear pulse part with time-variable parameters, and several non-linear elements, widens the range of problems which can be solved and, in particular, makes it possible to investigate non-linear pulse control systems in which pulse-width, pulse-phase and pulse-frequency modulation is provided.

### References

- 1 BROMBERG, P. V. *Stability and self-oscillations of sampled-data control systems*. 1953. Moscow
- 2 HAHN, W. Über die Anwendung der Methode von Liapunov auf Differenzengleichungen. *Math. Ann.* 136 (1958) 430-441

- <sup>3</sup> HAHN, W. *Theorie und Anwendung der direkten Methode von Liapunov*. 1959. Springer Verlag
- <sup>4</sup> KALMAN, R. E. and BERTRAM, J. E. Control system analysis and design via the second method of Liapunov: II. Discrete Time Systems. *Trans. ASME, J. of Basic Engineering*, S.D., 82 No. 3 (1960) 371-400
- <sup>5</sup> BERTRAM, J. E. *The direct method of Liapunov in the analysis and design of discrete time control systems*. Work session in Liapunov's second method. Ed. L. K. Kazda, University of Michigan (1960) 79-104
- <sup>6</sup> POPOV, V. M. Criterii de stabilitate pentru sistemele nelinare de reglaje automata po utilizarea transformatbazate pe Laplace. *Studii și Cercetări de Energetica*. An. 9, 1 (1959) 119-135
- <sup>7</sup> POPOV, V. M. Concerning the absolute stability of non-linear control systems (Ob absolyutnoi ustoychivosti nelineynykh sistem avtomaticheskogo regulirovaniya). *Avtomat. i telemekh* 22, No. 8 (1961) 961-979
- <sup>8</sup> LETOV, A. M. *The stability of non-linear control systems*. 1955. Moscow; Gostekhizdat
- <sup>9</sup> TSYPKIN, Ya. Z. *Theory of pulse systems*. 1958. Moscow; Fizmatgiz; English translation: *Sampling Systems and their Application*. 1962. London, Oxford; Pergamon Press
- <sup>10</sup> AKHIEZER, N. I. *The classical problem of moments*. 1961. Moscow; Fizmatgiz
- <sup>11</sup> AIZERMAN, M. A. Concerning a problem which touches on the stability 'in the large' of dynamic systems. *Uspekhi mat. nauk* 4, 4 (32) (1949) 186-188
- <sup>12</sup> TSYPKIN, Ya. Z. Periodic modes in non-linear pulse-type control systems. *Trud. Tashkent Polytechnical Institute (New series) Energetika*, 20 (1961) 184-195
- <sup>13</sup> TSYPKIN, Ya. Z. Elements of the theory of numerical automatic systems. *Automatic and Remote Control*. 1961. London; Butterworths. *Akad. Nauk SSSR*, 2 (1961) 63-79
- <sup>14</sup> POPOV, V. M. A criterion of the quality of non-linear control systems. *Automatic and Remote Control*. 1961. London; Butterworths; *Akad. Nauk SSSR* 1 (1961) 404-441
- <sup>15</sup> HOROWITZ, I. M. The sensitivity problem in pulse feedback systems. *IRE Trans. Automatic Control*, AC-6, No. 3 (1961) 251-260

## DISCUSSION

G. P. SZEGÖ, *Research Institute of Advanced Studies, 7212, Bellona Ave., Baltimore 12, MD., U.S.A.*

This paper is another important contribution that Professor Tsyppkin has made to the theory of sampled-data systems.

My comments range from the use of Liapunov functions for the solution of the problem treated in the first part of this paper, to the problem of absolute stability of sampled-data control systems. In the paper of Kalman<sup>1</sup>, published at the beginning of this year and in which he extends a result due to Yakubovich<sup>2</sup>, the connections between the second method of Liapunov and the method of Popov are shown for continuous systems, I have been concerned with the analogous problem for discrete systems.

I shall very briefly state the results obtained and outline the method used.

Given a sampled-data control system  $\Sigma$

$$x_{t+1} = Ax_t - a\varphi(\sigma_t) \quad (1)$$

$$\sigma_t = 2b'x_t \quad t=0, 1, \dots$$

$$\varphi(0)=0, \quad 0 < \sigma\varphi(\sigma) < \sigma^2 k \quad (2)$$

Assume  $A$  stable, with at the most one eigenvalue equal to unity, the linear part of  $\Sigma$  completely controllable and observable, and all linear systems of the class  $\varphi = \sigma m$ ,  $0 < m \leq k$ , asymptotically stable.

(a) It can be proved<sup>3</sup> that the stability condition (23) of the paper is a sufficient condition for the existence of a quadratic Liapunov function

$$v = x'Hx \quad (3)$$

which proves the absolute stability of  $\Sigma$ .

Consider the  $v$ -difference along the solutions of (1):

$$\Delta v = -[\gamma\varphi + q'x]^2 - \varphi \left[ \sigma_t - \frac{1}{k} \varphi \right] \quad (4)$$

where

$$(I) \quad H + A'HA = qq'$$

$$(II) \quad A'Ha - b = \gamma q \quad (5)$$

$$(III) \quad \frac{1}{k} - a'Ha = \gamma^2$$

For the existence of real  $\gamma$ ,  $q$  and  $H = H'$  satisfying (5) it is necessary and sufficient that

$$\frac{1}{k} + 2 \operatorname{Re} b'(Iz - A)^{-1}a \geq 0 \quad (6)$$

is satisfied for all real  $z = e^{i\omega}$ ,  $\omega$  real.

This condition may be rewritten as

$$\frac{1}{k} + \operatorname{Re} W(z) \geq 0 \quad (7)$$

where  $W(z) = 2b(Iz - A)^{-1}a$  is the open loop transfer function of the linear part of  $\Sigma$ . This condition is (23) of the paper. By the hypothesis of complete observability it follows that (4) is negative definite along (1). Thus (1) is absolutely stable if (6) is satisfied.

(b) Let us proceed further along this line. Assume that  $\varphi$  has continuous and bounded derivative and define

$$0 < \frac{d\varphi}{d\sigma} < k_0 \quad (8)$$

as in Figure 3 of the paper.

By using the Liapunov function

$$v = x'Hx + \int_0^\sigma \varphi(s) ds \quad (9)$$

one can show<sup>4</sup> that for (1) to be absolutely stable it is sufficient that

$$(\gamma + \beta) \frac{1}{k} + \operatorname{Re}(\gamma + \beta z) W(z) - \frac{k_0 |\beta|}{2} |(z-1) W(z)|^2 \geq 0 \quad (10)$$

$$(\gamma + \beta) \frac{1}{k} + 2[\beta a'b - k_0 |\beta| (a'b)^2] \geq 0, \gamma + \beta \geq 0$$

are satisfied for all  $z = e^{i\omega}$ ,  $\omega$  real and some real  $\gamma$  and  $\beta$ .

(c) Under the same assumption which led to the condition (26) and the relaxed conditions on  $A$ , such as the sixth section of the paper, by using the Liapunov function (7) one can show<sup>5</sup> that the system (1) is absolutely stable if

$$\begin{aligned}
 (\alpha - \beta) \frac{1}{k} + \operatorname{Re}(\beta z - \alpha) \frac{W(z)}{1 + k_0 W(z)} &\geq 0 \\
 2\beta a'b + (\alpha - \beta) \frac{1}{k}, \alpha - \beta > 0, \beta > 0
 \end{aligned}
 \quad (11)$$

are satisfied for all  $z = e^{i\omega}$ ,  $\omega$  real and some real  $\gamma$  and  $\beta$ .

Even if we have now some fairly good results on this problem the status of stability theory for sampled-data systems has still not reached a satisfactory stage as in the continuous case. Further work is needed and improvements can be achieved by using the new method of Popov<sup>5,6</sup> and very likely by using some more sophisticated Liapunov functions.

Professor Tsytkin mentioned in the paper the possibility of expressing the basic frequency criteria of absolute stability in analytic form. Could he discuss this point in some detail?

## References

- <sup>1</sup> KALMAN, R. E. *Proc. Nat. Acad. Sci., Wash.* 49 (1963) 201–205
- <sup>2</sup> YAKUBOVICH, V. A. *Dokl. Akad. Nauk SSSR*, 143 (1962) 1304–7
- <sup>3</sup> SZEGÖ, G. P. and KALMAN, R. E. *C.R. Acad. Sci., Paris*
- <sup>4</sup> SZEGÖ, G. P. *Proc. Nat. Acad. Sci., Wash.* 49 (1963)
- <sup>5</sup> SZEGÖ, G. P. *C.R. Acad. Sci., Paris*, in the Press
- <sup>6</sup> POPOV, V. M. *C.R. Acad. Sci., Paris*, 256 (1963) 3568
- <sup>7</sup> HALANAY, A. *C.R. Acad. Sci., Paris*, 256 (1953) 4818

E. I. JURY, *Department of Electrical Engineering, University of California, Berkeley, California, U.S.A.*

This paper represents a useful and important contribution to the absolute stability of non-linear discrete (pulse) control systems. The main feature of the method of this paper is the extension of Popov's approach (applied to the continuous case) to the discrete case, thus simplifying the underlying procedure for stability test and assessing the quality of control.

While the method of this paper mainly provides a sufficient condition for stability, it also provides for certain non-linear discrete systems the necessary and sufficient condition as explained by the author. This extended feature, among others, provides an added advantage on the use of Liapunov's quadratic form. For this class of systems in which were obtained the necessary and sufficient condition, I might mention that *Figure A* of this discussion could be added to the forms indicated by Professor Tsytkin in *Figure 7(a)* and *(b)* of his paper.

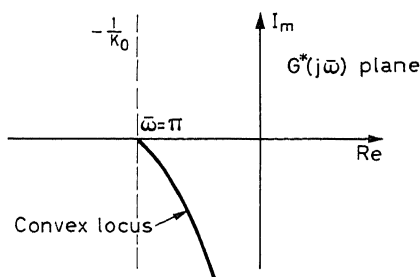


Figure A

More than a decade ago, I showed in my Doctoral Thesis that such a convex locus could represent the limiting case when  $T$  (the sampling period) is very large compared with the linear system time constants. It would be interesting to prove analytically that for this limiting case, Tsytkin's method yields the necessary and sufficient condition for stability.

The class of system where the method yields a necessary and sufficient condition is rather limited (confined to first-order or the limiting case) and thus for the most encountered systems the method represents the sufficient condition. In view of the latter restriction it is of interest to discuss the problem of possible improvements on the sufficiency

condition. This problem which was stimulated by private correspondence with the author, has been recently studied and certain results have been obtained which are briefly discussed in the following.

In order to improve on the sufficiency condition, I find, in collaboration with Mr. B. W. Lee, that this could only be done if some other restrictions on the non-linearity discussed in Tsytkin's equation (1) is further imposed. This additional restriction limits the so-called (a.c.) gain\* of the non-linear element to less than  $K'$ . For such a class of systems the following theorem has been proposed and the proofs will be published in the near future. The stability theorem requires that the following relationship be satisfied on the unit circle, i.e.

$$\operatorname{Re} G^*(z) [1 + q(z-1)] + \frac{1}{K} - \frac{K'q}{2} |(z-1)G^*(z)|^2 \geq 0$$

where  $q$  is a non-negative number and  $\operatorname{Re}$  represents the real part.

Certain features of the stability criterion embodied in the above theorem can be briefly mentioned.

(1) The application of the criterion requires knowledge only of the  $Z$  transform of the linear plant on the unit circle ( $z = \exp.j\omega$ ), which may be derived easily from the frequency response of the linear plant and the compensating network. This is a characteristic feature of results obtained by Popov's method.

(2) This criterion is directly applicable to systems with linear plants of any order. This follows since the criterion is stated in terms of the  $Z$  transform of the linear plant rather than the time-domain state variables of the system.

(3) The criterion includes Tsytkin's results as a special case, i.e. when  $q = 0$ .

(4) For the limiting case when  $T \rightarrow 0$  (i.e. the discrete system approaches the continuous case) this criterion can be shown to coincide with Popov's method.

For the brevity of the discussion, we will not elaborate on the above points since these will be discussed in detail in a future publication; however, at the conclusion one may note the following:

A general comparison between Tsytkin's criterion and the one presented above is somewhat difficult since the class of systems considered by the former is more general. This is due to the extra restriction we have imposed on  $K'$ . However, for all cases where least conservative results are obtained with  $q = 0$ , results obtained by either criterion may be evaluated on an identical basis. If  $q \neq 0$ , results are comparable only when the system is restricted to being a member of the class considered in the new criterion. This method yielded better results, for the examples worked out for this class of systems, than those obtained with the method mentioned in the paper.

## Reference

- JURY, E. I. and LEE, B. W. On the stability of a certain class of non-linear sampled-data systems. *I.E.E.E. Transactions on Automatic Control*. In the Press

C. C. LI, *Department of Electrical Engineering, University of Pittsburgh, Pittsburgh, Pennsylvania, U.S.A.*

The author is commended for this excellent contribution to the non-linear pulse control system theory. A research project on determining the discrete system stability and minimizing the sum of quadratic error, via Liapunov's second method, has been done by our group at the University of Pittsburgh. The author may be interested in noting R. P. O'Shea's Sc. D. Thesis 'On the Determination of Domain of Asymptotic Stability of a Class of Non-linear Discrete-Data Control Systems', Department of Electrical Engineering, University of Pittsburgh, 1962.

Has the author actually applied this method to pulse frequency modulated control systems?

\* The slope or derivative of the non-linearity.

CHING-SHUAN-SHUE, *Harbin Institute of Technology, China*

Up to the present time, theory of non-linear pulse control systems has dealt mostly with the investigation of periodic oscillations, which, however, are not the frequently occasioning conditions of non-linear pulse control systems.

This interesting paper by Professor Tsyarkin gives some new means of studying the non-linear pulse control systems. It offers a frequency criterion of absolute stability by the concepts of the linear theory of pulse systems and a method of estimation of indices of the quality of processes.

However, there are two points in the article that are not very clear. If the parameters of the linear pulse part, or those of the non-linear element, are time-varying, for example, when they are periodic functions of time, will the frequency criteria of stability remain unchanged or would some corrections have to be made?

Moreover, this article considers the cases, where the lattice function  $f[n]$  approaches zero as limit when  $n$  approaches infinity. Will the proposed method remain unchanged when

$$\lim_{n \rightarrow \infty} f[n] = \text{const.}$$

YA. Z. TSYPKIN, *in reply*

I am grateful to Dr. G. P. SZEGÖ, Professors E. I. Jury, C. C. Li and Ching-shuan Shue for their very interesting discussions and some not less interesting questions and remarks.

Dr. Szegö has pointed out the relationship between the criterion of absolute stability proposed with the second method of Liapunov and states that inequality (23) is sufficient for a quadratic function of Liapunov. This fact was well known to us. The further development of absolute stability criteria under the assumption of the additional limitation

$$0 \leq \frac{d\varphi(\sigma)}{d\sigma} < K_0$$

is most interesting. On the other hand, I cannot agree with the opinion of Dr. Szegö that the problem of absolute stability of non-linear pulse systems does not reach a satisfactory level as in the case of analogue

systems. On the contrary, the solution of this problem for non-linear pulse systems has a more simple and accomplished form. As concerns the formulation of analytical conditions for absolute stability, only the preliminary results were obtained in this direction and they are unfortunately given in a very complicated form. Professor Jury confirms that the criterion of absolute stability proposed, very often gives extremely hard conditions. Using the additional limitations

$$\left(0 \leq \frac{d\varphi}{dx} < K'\right)$$

these conditions can be eased. The inequality given by Professor Jury most probably coincides with the inequality proposed by Dr. Szegö. It must be pointed out that these inequalities give the solution of the problem of absolute stability in contrast to the problem of our paper considered under the existence of two limitations. Though the solution of this problem has a more complicated form, it is nevertheless of great practical interest as it gives a possibility of obtaining wider stability conditions.

I think there is no need to add to *Figure 7(a)* and *(b)* of my paper the case shown by Professor Jury in his *Figure A*, as much as it can be considered as a particular case of *Figure 7(b)* when the frequency limit is  $\bar{\omega}_{\text{lim}} = \pi$ .

Professor Li naturally is interested in investigations of absolute stability of pulse frequency systems. The criterion we proposed can be applied directly to the pulse frequency systems of the first order, but by the corresponding generalization given, it can be applied to the systems of higher order as well.

I am grateful to Professor Li for mentioning the paper which concerns the question of investigation of stability of non-linear pulse systems. It will be most interesting to read this.

In reply to Professor Ching-shuan Shue's questions I would like to say that the criterion's proof itself clearly shows that it holds, even when non-linear unit characteristics and sampler gain are arbitrarily varying. In other cases the criterion should be altered.

If  $\lim_{n \rightarrow \infty} f[n] = \text{const.}$ , then one cannot use the asymptotic stability concept for  $x = 0$ , but must solve Liapunov's problem or investigate process stability.

The criterion is applicable even in this case.

# Synthesis of Optimum Sampled-data Systems

L. N. VOLGIN

## Summary

In this paper the problems connected with the synthesis of optimum pulse systems for the control of dynamic media by means of linear models are considered. These problems are stated, using the mathematical tool of polynomial equations, which make possible the synthesis of efficient pulse systems. The analytical conditions for the efficiency of the pulse systems, which combine the usual conditions of stability of linear systems with the conditions of 'approximation' of modelling, are formulated. The inadequacy of the usual criteria of stability for the synthesis of efficient systems is demonstrated. An outline is given of the possible applications of the method of polynomial equations in the theory of automatic control.

## Sommaire

Ce rapport traite des problèmes concernant la synthèse de systèmes optimaux à impulsions pour la commande d'éléments dynamiques au moyen de modèles linéaires. La solution de ces problèmes est obtenue en faisant usage des équations polynomiales qui permettent la synthèse des systèmes à impulsions. Les conditions analytiques qui donnent la possibilité d'apprécier l'efficacité de ces systèmes en combinant les conditions usuelles de stabilité des systèmes linéaires avec les conditions de «l'approximation» de la simulation sont mises en évidence; on démontre l'insuffisance des critères de stabilité usuels pour la synthèse de tels systèmes. La possibilité de l'application de la méthode des équations polynomiales dans la théorie de réglage automatique, est esquissée.

## Zusammenfassung

In diesem Beitrag wird die Synthese von optimalen Abtastsystemen für die Regelung dynamischer Strecken mittels linearer Modelle behandelt. Diese Probleme werden mathematisch durch Polynomgleichungen formuliert, wodurch der Entwurf leistungsfähiger Abtastsysteme möglich wird. Die analytischen Bedingungen für die Leistungsfähigkeit von Abtastsystemen, welche die üblichen Stabilitätsbedingungen linearer Systeme ebenso umfassen wie die Bedingungen für „Näherung“ durch Modell, werden ausgearbeitet. Es wird nachgewiesen, daß die üblichen Stabilitätskriterien für den Entwurf leistungsfähiger Systeme ungenügend sind. Die regelungstheoretischen Anwendungsmöglichkeiten für die Methode der polynomen Gleichungen werden gezeigt.

## Introduction

The transition to the scientific design of compound automatic complexes and the resulting increase in the calculation difficulties demand the finding of new ways for the formalization of solutions and simplification of calculation. The creation of the methods of linear, non-linear and dynamic programming should be regarded as considerable achievements in this field. The method of polynomial equations used below, the efficiency of which was demonstrated on a number of problems of automatic control, may be added to this number of methods. The polynomial equations consist of a variety of Diophant equations, for the specific methods of solution of which are easily programmed for digital computers. The development of the operator method

of analysis for the linear pulse systems, of the discrete Laplace transformation, or  $z$ -transformation (see Tsyarkin<sup>1</sup>, Gurevich<sup>2</sup>, Zadeh and Ragazzini<sup>3</sup>, and others), and the emergence of a large number of different methods for the synthesis of the optimum linear pulse systems (the works of Tsyarkin<sup>4-6</sup>, Bergen and Ragazzini<sup>7</sup>, Chang<sup>8-10</sup>, Jury<sup>11</sup>, Bertram<sup>12</sup>, Potapov<sup>13</sup>, Krasovskii<sup>14</sup>, Perov<sup>15</sup>, and others) were the reason for the creation of the method of polynomial equations.

At present the theory of optimum pulse systems for the control of linear plants lies at the foundation of design of self-optimizing systems, which contain digital computers. In these systems the automatic linearization of equations for the plant during the operation of the system is achieved on the basis of principles described in the works of Kalman<sup>16</sup>, Bigelow and Ruge<sup>17</sup>, and others. Thus, the theory of optimum linear pulse systems develops into the theory of an extensive class of self-optimizing pulse systems of control, adaptable to the changing characteristics of the controlled plant and to the parameters of external signals.

The basic difficulties which arise in the design of the systems containing an optimizing model for the plant are associated with the violation of the conditions of 'approximation' of simulation, which require a continuous relationship between the quality of control and the change in the parameters of the plant being simulated. Some of the above-mentioned authors touch upon the questions of control for the plants with negative dynamic properties, during the compensation of which the violation of 'approximation' is possible. The criteria of approximation found under these conditions, which do not allow the contraction of the zero and poles of the transfer function of the plant for the individual structures of automatic systems, served as the starting point for the search of analytical conditions of approximation, suitable for any structures. The investigation of the conditions of approximation for automatic systems showed that they are closely connected with the conditions of stability, and that the distinctive feature of these conditions is based on the distinctive ideas about the 'coordinate' and 'parameter'. The conditions found below, which combine the conditions of stability and approximation, are called *the efficiency conditions*, since the term 'efficiency' literally reflects the essence of the considered phenomenon. From the analytical conditions of efficiency for different structures of automatic systems emerge different criteria for efficiency. On the basis of these criteria it is possible to conclude that the criteria of stability adopted at present are inadequate for the synthesis of efficient systems. The attempts to solve the problems of synthesis for automatic systems often encountered in literature, inaccurate on the whole or having a very limited field of application, are explained by this. The author has shown<sup>18-20</sup> that the polynomial equations represent a mathematical tool which is adequate for the problem of synthesis for efficient automatic pulse systems. A systematic treatment of the method of

polynomial equations is contained in the author's monograph<sup>20</sup>. In the given paper a derivation of the analytical conditions of efficiency is given, and a brief survey made of the problems of automatic control, solvable by means of polynomial equations.

### Denotations and Terms Used

1. Symbol  $z$  is used for transformations, where  $z$  is the delay operator for a single cycle.
2. The systems and signals are represented by the rational real functions of  $z$  of the form  $F = A/B$ , where  $A$  and  $B$  are polynomials of  $z$ .
3. The factorization of functions  $F$  with reference to contour  $\Gamma\{|z| = 1\}$  gives the real functions  $F^+$  and  $F^-$ ;  $F = F^+F^-$ , where the sign  $F^+$  denotes the absence of zeros and poles of the function in the region  $D^-\{|z| \leq 1\}$ , and the sign  $F^-$  denotes their absence in the region  $D^+\{|z| > 1\}$ .
4. The separation of functions  $F$  with respect to contour  $\Gamma$  gives the real functions  $F = F_+ + F_-$ , where the sign  $F_+$  denotes the absence of poles of the function in the region  $D^-$ , and the sign  $F_-$  denotes their absence in the region  $D^+$ .
5. The representation (transfer function) of the controlled plants will be made by  $G = P/Q$ , where  $P$  and  $Q$  are polynomials of  $z$ ; the representation (programme) of the pulse unit will be made by  $W = C/D$ , where  $C$  and  $D$  are polynomials of  $z$ ; the representation of the pulse system as a whole will be made by  $H$ , and the representations for the input and output are equal to  $X$  and  $Y$  respectively.

### Analytical Conditions for the Efficiency of Pulse Systems

By considering the mathematical model of an actual physical system, one is deliberately making a differentiation between the 'coordinates' of the system, the changes in which are reflected by the given model, and its 'parameters' which are determined as fixed numbers which, in the given model, form the basis for calculations. However, the practice of construction of automatic systems shows that the uncontrollable discrepancies between the calculated and the actual parameters may be the cause of profound disparity in the calculated and the actual behaviour of the system. The failure to take this fact into account will sometimes lead to the construction of inefficient systems. The majority of automatic systems (the systems of stabilization and programme control, the computer and the reproduction systems, the systems for transmission and processing of data) require a continuous relationship for the behaviour and the small changes in external conditions, which are expressed in the change of input coordinates and parameters of the system. The conditions for which a continuous relationship between the coordinates of the system is observed are the conditions of stability. The conditions for which a continuous relationship between the behaviour of the system and the deviations of its parameters from the calculated values, which are assumed to be constant in a given model, is observed, are the conditions of approximation of simulation. The general condition of efficiency for an automatic system, constructed on the basis of a definite calculated model, which unites the conditions mentioned, may be formulated as follows. *With small variations in the input coordinates and parameters of the system the variations in the output coordinates should be small.*

Let us find the analytical conditions for the efficiency of an automatic pulse system, with a single input and a single output coordinate, described by the following difference equation:

$$\mathcal{F}(x_i, x_{i-1}, \dots, x_{i-n}, y_i, y_{i-1}, \dots, y_{i-m}) = 0 \quad (1)$$

where  $\mathcal{F}$  is the continuous function differentiable with respect to all arguments,  $i$  is the discrete time, and  $n$  and  $m$  are the corresponding numbers of stored values  $x$  at the input and  $y$  at the output. At the foundation of calculation of the system lies the linear model, obtainable by means of linearization of the equation of the system in the vicinity of the 'operating point':

$$\sum_{k=0}^n \left( \frac{\partial \mathcal{F}}{\partial x_{i-k}} \right)_0 \delta x_{i-k} + \sum_{k=0}^m \left( \frac{\partial \mathcal{F}}{\partial y_{i-k}} \right)_0 \delta y_{i-k} = 0 \quad (2)$$

The numbers

$$a_k = \left( \frac{\partial \mathcal{F}}{\partial x_{i-k}} \right)_0; \quad b_k = - \left( \frac{\partial \mathcal{F}}{\partial y_{i-k}} \right)_0 \quad (3)$$

which do not depend on index  $i$  over the interval of time under consideration, represent the equivalent parameters of the linear model.

Using  $z$ -transformation of number sequences<sup>20</sup>, the equation for the linear model (2) may be written in the form:

$$Y = HX \quad (4)$$

where  $H$  is the representation of the model, which is the rational function

$$H = \frac{A}{B}, \quad A = \sum_{k=0}^n a_k z^k, \quad B = \sum_{k=0}^m b_k z^k \quad (5)$$

The representation of a real system, the parameters of which change in relation to time and coordinates, but sometimes also in an unexpected form, differs from the representation of its model by the variations  $\delta H, \delta^2 H, \delta^3 H, \dots$ , which must satisfy the general condition for the efficiency of the system.

By varying the relation (4), the corresponding variations for the output of the system are obtained:

$$\begin{aligned} \delta Y &= H \cdot \delta X + \delta H \cdot X \\ \delta^2 Y &= H \cdot \delta^2 X + 2 \delta H \cdot \delta X + \delta^2 H \cdot X \end{aligned} \quad (6)$$

The conditions under which the variations in the output coordinate remain small have the form:

$$(\delta Y)_- = 0; (\delta^2 Y)_- = 0; (\delta^3 Y)_- = 0, \dots \quad (7)$$

By separating the right sides of expressions (6) the analytical conditions for the efficiency of the pulse system are obtained:

$$H_- = 0; (\delta H)_- = 0; (\delta^2 H)_- = 0, \dots \quad (8)$$

in which case the first of these conditions is the usual condition of stability, whereas the last are the conditions of 'approximation' of simulation. The necessity for taking into account the large variations is caused by the fact that as regards the parameters of the system its representation is a non-linear function. It is possible to construct an example where the violation of the efficiency is caused as much as is desired by a high variation<sup>20</sup>. However, in practice, mostly violations of the first two conditions of efficiency are encountered.



### Criteria for the Efficiency of the Basic Structures of Automatic Pulse Systems

The method of combining the controlled plants and the computing units is called the structural system of control. The simplest pulse systems of automatic control contain a single computing unit with representation (programme)  $W$  and a single controlled plant with representation  $G$ . To each structure of the system of control corresponds a definite function  $H$ , which depends rationally on  $W$  and  $G$ :

$$H = H(W, G) \quad (9)$$

which is called the representation of the system. For each structure of control there is a definite class of permissible functions  $H$ , which may be realized in the system by the choice of different control programmes  $W$ , remaining at the same time within the limits of conditions of efficiency. The structures, which permit the realization of arbitrary functions  $H$  are called *ideal structures*. The structures which do not have even a single permissible function are called *inefficient structures*. From the point of view of the condition of stability only the stable functions of type  $H_+$  are permissible functions. However, if it is necessary to realize an unstable function, then the condition of stability may be discarded by limiting oneself to the fulfilment of the conditions of approximation.

By taking into account the variations in the representation of the controlled plant, simulated by function  $G$ , the conditions of efficiency (8) applied to system (9) may be written in the form:

$$H_- = 0; \left( \frac{\partial H}{\partial G} \cdot \delta G \right)_- = 0; \left( \frac{\partial^2 H}{\partial G^2} \cdot \delta^2 G \right)_- = 0; \dots \quad (10)$$

The functions  $H$ ,  $\partial H / \partial G$ ,  $\partial^2 H / \partial G^2$ , ..., derived by differentiation of (9), depend on  $W$  and  $G$ . In synthesis of systems for the automatic control of the programme of the computing unit,  $W$  is chosen in relation to the representation of plant  $G$ :

$$W = W(G) \quad (11)$$

The verification of the synthesized systems for efficiency is made by the substitution of this relationship in the expression (10) after carrying out the operations of differentiation in them.

In a general case the pulse systems of automatic control contain several controlled plants and computing units, which are connected up into a single structure. These systems may have several inputs and outputs. The verification of the conditions for efficiency should be carried out in this case by the variation of all the output coordinates for the variation in the representations of all the controlled plant.

The compensation for the negative dynamic properties of the controlled plant, by means of the computing unit having the same negative dynamic properties, is the cause of violation of the conditions of efficiency of pulse systems of automatic control. Namely, such a compensation takes place, for example, during the trivial recalculation of the programme for the computing unit  $W$  for a simple closed system, the representation of which is:

$$H = \frac{WG}{1 + WG} \quad (12)$$

by the formula:

$$W = \frac{1}{G} \cdot \frac{H}{1 - H} \quad (13)$$

by proceeding from the initial function  $H$ , which is chosen without taking into account the conditions of efficiency.

This assumption will be proved. By carrying out the factorization of the representation of the plant it is obtained that:

$$G = G^+ G^- \quad (14)$$

Functions  $G^+$  and  $G^-$ , equal to:

$$G^+ = P^+ / Q^+; \quad G^- = P^- / Q^- \quad (15)$$

are the positive and the negative portions of the representation of the plant.

The *positive* plant, which has a representation  $G^+$ , is characterized by the following dynamic properties: stability, instantaneousness of reaction, and smoothness of the transition process. The *negative* plant, which has a representation  $G^-$ , displays negative dynamic properties: instability, retardation of reaction, and sudden ejections in the transition process.

By modifying formula (12) one obtains:

$$\delta H = \frac{W}{(1 + WG)^2} \cdot \delta G; \quad \delta^2 H = -\frac{2W^2}{(1 + WG)^3} \cdot \delta^2 G; \dots \quad (16)$$

First of all, conditions will be found under which the closed system is ideal, i.e. capable of reproducing the arbitrary function  $H$ . The corresponding programme for the computing unit is chosen in accordance with formula (13). By substituting this formula in (16) one obtains:

$$\delta H = H(1 - H) \frac{\partial G}{G}; \quad \delta^2 H = -2H^2(1 - H) \frac{\partial^2 G}{G^2}; \dots$$

$$\text{or} \quad \delta H = H(1 - H) \left( \frac{\delta P}{P} - \frac{\delta Q}{Q} \right);$$

$$\delta^2 H = -2H^2(1 - H) \left( \frac{Q\delta^2 P}{P^2} - \frac{2\delta P \cdot \delta Q}{P^2} + \frac{2\delta^2 Q}{PQ} \right); \dots \quad (17)$$

The conditions of efficiency (8) require that  $P^- = Q^- = 1$ . Thus, the closed system is ideal only in that case when the plant is positive. In the case of the plant with negative dynamic properties the function  $H$  is not realizable because of the violation of the conditions of approximation.

It will be shown that the closed automatic system is efficient for any controlled plant, under which conditions the class of permissible functions of this system is equal to

$$H = P^- \theta F_+ \quad (18)$$

where  $F_+$  is the arbitrary stable rational function of the form:

$$F_+ = A/B^+ \quad (19)$$

and  $\theta$  is the polynomial which satisfies the polynomial equation in respect of the unknown polynomials  $\theta$  and  $\Pi$ :

$$AP^- \theta + Q^- \Pi = B^+ \quad (20)$$

The corresponding programme of control has the form:

$$W = \frac{AQ^+ \theta}{P + \Pi} \quad (21)$$

It will be verified whether the conditions of efficiency are fulfilled. By substituting (21) in (16) and by taking into account (20) one obtains:

$$\delta H = \frac{A\theta\Pi}{(B^+)^2} \cdot \frac{Q\delta P - P\delta Q}{P^+Q^+};$$

$$\delta^2 H = \frac{2A^2\theta^2\Pi}{(B^+)^3} \cdot \frac{Q^2\delta^2 P - 2Q\delta P\delta Q + 2P\delta^2 Q}{(P^+)^2Q^+}; \dots$$

The conditions of efficiency are fulfilled for any values of  $G$ . In the case of a stable controlled plant the polynomial  $\theta$ , as follows from the polynomial eqn (20), becomes arbitrary, and the class of permissible functions is extended to

$$H = P^- F_+ \quad (22)$$

Thus, one proves the criterion for the efficiency of a closed system, which requires in addition to the fulfilment of the usual criterion of stability, that the programme of the computing unit does not shorten polynomials  $P^-$  and  $Q^-$ .

Using the analytical conditions of efficiency, it is possible to derive the criterion of efficiency for any structures of automatic systems. By means of these conditions it is easy to prove, for example, the following well-known propositions:

- (1) The systems on the limit of stability are inefficient.
- (2) The open systems of control are efficient only for the stable plants.
- (3) The ideal structures of control for plants having negative dynamic properties do not exist.
- (4) The parallel system of control is ideal for stable plants; the sequential (cascade) system of control is ideal only for positive plants.

In view of the non-existence of ideal structures of control for the arbitrary plants, the criterion of efficiency of the automatic system is more rigid than the criterion of stability. Only for positive plants are the general criteria of stability of the linear systems adequate.

In order not to violate the conditions of efficiency, the optimum function  $H$  of the system should be sought for in the class of permissible functions. The wider the class of permissible functions for a given structure, the higher the quality of the optimum system, remaining conditions being equal. Therefore, in the synthesis of a system of control for a given plant, a structure of control having as wide a class of permissible functions as possible, a structure close to an ideal one, should be chosen.

### The Use of Polynomial Equations in the Synthesis of Optimum Pulse Systems

It has been established that the classes of permissible functions for pulse systems are expressed in terms of polynomial equations. In the author's work<sup>18-20</sup>, it was shown that the synthesis of optimum pulse systems of control for linear plants based on a number of basic criteria may be made entirely

\* Applicable to the stable plants the criterion of non-contraction  $P^-$  was, for the first time, introduced in the work of Bergen and Ragazzini<sup>7</sup>.

by means of polynomial equations. The finding of the optimum programme of control is, as a rule, reduced to the solution of a system of polynomial equations. The computation methods for the solution of a system of polynomial equations applicable to the use of digital computers have also been developed and their advantage over the ordinary methods in the synthesis of controlled programmes for plants of a high order with complex correlational relationships was proved. By means of the polynomial equations, a number of new problems of automatic control, in particular for the unstable controlled plants, was solved. The basic problems for the synthesis of pulse systems and their solutions, obtained by the method of polynomial equations, omitting the proofs because of the lack of space, are now enumerated.

The problem of synthesis of the pulse system with the minimum transient period for a given input action:

$$X = R/S \quad (23)$$

where  $R$  and  $S$  are the polynomials of  $z$ , is reduced to the solution of the following polynomial equation:

$$P^-\theta + SQ^-\Pi = R \quad (24)$$

in respect of unknown polynomials  $\theta$  and  $\Pi$ . The corresponding controlling programme is equal to:

$$W = \frac{Q^+\theta}{P^+S\Pi} \quad (25)$$

The representation of the transient process has the form:

$$E = Q^-\Pi \quad (26)$$

The minimum duration of the transient process, which ensures the fulfilment of the conditions of efficiency, from the number of cycles, is equal to the sum of powers of polynomials  $P^-$  and  $Q^-$ .

With the limitation for the module of the controlling action:

$$|u_i| \leq r \quad (i=0, 1, 2, \dots) \quad (27)$$

the corresponding problem is reduced to the finding of a non-minimum solution of the polynomial equation, which is found by special computing methods. The modification of the polynomial equation (24) leads to the derivation of a system which has no hidden oscillations.

The problem of synthesis based on the criterion of the minimum of the total quadratic error:

$$\mathcal{J} = \sum_{i=0}^{\infty} e_i^2 = \frac{1}{2\pi j} \oint_r E(z) E(z^{-1}) \frac{dz}{z} \quad (28)$$

is reduced to the solution of the system consisting of two polynomial equations:

$$\left. \begin{aligned} P^-\theta + Q^-\Pi &= I^+ \tilde{P}^- \tilde{Q}^- \\ P^-\theta + U^+\phi &= I^+ \tilde{P}^- \tilde{Q}^- \end{aligned} \right\}^* \quad (29)$$

in respect of the unknown polynomials  $\theta$ ,  $\Pi$  and  $\phi$ . The polynomials  $I$  and  $U$  are the numerator and denominator of

\* Polynomial with the reversed order for the sequence of coefficients is denoted by symbol  $\tilde{A}$ .

function  $X(z)X(z^{-1})$ . The corresponding controlling programme is equal to

$$W = \frac{Q + \theta}{P^+ \Pi} \quad (30)$$

The calculation of the quadratic error may also be made by means of the polynomial equation<sup>20</sup>.

The problems of synthesis of the optimum pulse systems of automatic control and of processing of data for the random input signals, by taking into account the universal nature and the prevalence of quadratic dispersion criteria, represent the most favourable field for the application of polynomial equations. The general problem of synthesis of a pulse system, optimum according to the criterion of dispersion of the error for finite time of transition into the unshifted state, is reduced to the solution of a system consisting of three polynomial equations, one of which secures the efficiency of the synthesized system, the second, the finiteness of the settling time, and the third, the minimization of dispersion of the error. The solution of this general problem determines the solutions of the numerous particular problems of extrapolation, filtration, differentiation and integration of random processes by means of pulse computing units. The optimization of the pulse systems, by arbitrary criteria of quality, is reduced to the combination of the method of polynomial equations and the general methods of mathematical programming. By means of the theory of polynomial equations it is possible to synthesize the most economic programmes for the processing of data by the method of least squares. The obtained results show that the polynomial equations represent a suitable mathematical tool for the programming of many procedures of computer mathematics and of mathematical statistics, which are widely used in the self-optimizing systems of automatic control.

## Conclusions

The conditions of efficiency, formulated in this paper, limit the possibility of change in the dynamic properties of controlled plants by means of pulse computing units. Under these conditions the worst properties of the plant—instability, retardation, fluctuation—are shown to be the most difficult to overcome. The limits of the accuracy of control for dynamic plants by means of the pulse computing units whilst being wider than for the units of the continuous type, are, however, not limitless. Physically, this means that the inertia of the plants cannot be completely overcome. The problem of the theory of automatic control lies in the further clarification of the limits of possible accuracy of control, and the achievement of these possibilities

through the design of the most perfect controlling machines. It is hoped that the future development of polynomial equations will prove to be one of the important aids in the solution of this problem.

## References

- <sup>1</sup> TSYPKIN, Ya. Z. *Theory of Pulse Systems*. (Monogr.) 1958. Moscow; Fizmatgiz
- <sup>2</sup> JAMES, H., NICHOLS, N. and PHILLIPS, R. *Theory of servomechanisms*. Chap. V. (Monogr.). 1947. New York-London.
- <sup>3</sup> ZADEH, L. A. and RAGAZZINI, J. R. The analysis of sampled-data systems. *Trans. Amer. Inst. elect. Engrs* 71 Pt II (1952)
- <sup>4</sup> TSYPKIN, Ya. Z. Design of a system for automatic control under stationary actions. *Avtomat. Telemekh., Moscow* 4 (1953)
- <sup>5</sup> TSYPKIN, Ya. Z. Some questions relating to the synthesis of automatic pulse systems. *Avtomatika* 1 (1958)
- <sup>6</sup> TSYPKIN, Ya. Z. Optimum processes in automatic pulse systems. *Izv. AN SSSR, OTN, energet. avtomat.* 4 (1960)
- <sup>7</sup> BERGEN, A. R. and RAGAZZINI, J. R. Sample-data processing techniques for feedback control systems. *Trans. Amer. Inst. elect. Engrs* 73 Pt II (1954)
- <sup>8</sup> CHANG, S. S. L. Statistical design theory for strictly digital sampled-data systems. *Trans. Amer. Inst. elect. Engrs* 76 Pt I (1957)
- <sup>9</sup> CHANG, S. S. L. Statistical design theory for digital-controlled continuous systems. *Trans. Amer. Inst. elect. Engrs* 77 Pt II (1958)
- <sup>10</sup> CHANG, S. S. L. *Synthesis of Optimum Control Systems*. 1961. New York
- <sup>11</sup> JURY, E. I. *Sampled-data Control Systems* 1958. New York
- <sup>12</sup> BERTRAM, J. E. Factors in the design of digital controllers for sampled-data feedback systems. *Trans. Amer. Inst. elect. Engrs* 75 Pt II (1956)
- <sup>13</sup> POTAPOV, M. D. The problem of finite time of control and peculiarities of synthesis of some systems of automatic control. *Trudy VVIA im. N.E. Zhukovskogo*, 1959
- <sup>14</sup> KRASOVSKII, A. A. Synthesis of the correcting pulse units of the control systems. *Avtomat. Telemekh., Moscow* 6 (1959)
- <sup>15</sup> PEROV, V. P. *Statistical Synthesis of Pulse systems*. (Monogr.) Sovetskoe radio, 1959
- <sup>16</sup> KALMAN, R. E. Design of self-optimizing control systems. *Trans. Amer. Soc. mech. Engrs* 80, 2 (1958)
- <sup>17</sup> BIGELOW, S. C. and RUGE, H. An adaptive system using periodic estimation of the pulse transfer function. *I.R.E. Conv. Rec.* IV (1961)
- <sup>18</sup> VOLGIN, L. N. Method of synthesis of linear pulse systems for automatic control based on dynamic criteria. *Avtomat. Telemekh., Moscow* 20 No. 10 (1959)
- <sup>19</sup> VOLGIN, L. N. and SMOLYAR, L. I. The correction of control systems by means of certain calculating units. *Avtomat. Telemekh., Moscow* 21 No. 8 (1960)
- <sup>20</sup> VOLGIN, L. N. *The Fundamentals of the Theory of Controlling Machines*. (Monogr.) Sovetskoe Radio, 1962

# Integral Pulse Frequency Modulated Control Systems\*

C. C. LI and R. W. JONES

## Summary

The paper describes a new class of discrete data control systems in which a portion of the system makes use of pulse frequency modulation. This concept arose as an abstraction from the study of neural communication links in physiological control systems. An integral pulse frequency modulator is defined as one which emits a standard pulse whenever the integral of the input variable reaches a threshold value. The frequency of the output pulse train varies in a linear manner with the input magnitude. Modulators producing pulse trains of one or two signs are introduced and approximate frequency characteristics are obtained; they have discontinuous and discrete features, and depend upon signal amplitude, frequency, and initial phase angle. Feedback control systems making use of single-signed modulators in two parallel feedforward paths are described, and some of the unique stability considerations are studied. Continuous systems that approximate these stable discrete-data systems are constructed, and their transient properties considered.

## Sommaire

Le rapport décrit une nouvelle classe de systèmes de commande à données discrètes où une partie du système utilise la modulation de fréquence d'impulsions. Cette idée est une extrapolation de l'étude des liaisons de communications nerveuses des systèmes de commande physiologiques. Un modulateur intégral de fréquence d'impulsions (IPFM) est, par définition, tel qu'il émet une impulsion standard lorsque l'intégrale de la variable d'entrée atteint un seuil. La fréquence du train d'impulsions de sortie varie linéairement avec l'ampleur du signal d'entrée. Des modulateurs produisant des trains d'impulsions à un ou deux signes sont présentés et les caractéristiques approximatives de fréquence sont obtenues. Elles ont caractère discontinu et discret et dépendent de l'amplitude, de la fréquence et de l'angle de phase initial du signal d'entrée. Des systèmes de commande à réaction utilisant des modulateurs à un seul signe dans deux branches parallèles à chaîne ouverte sont décrits et certaines des considérations uniques de stabilité sont étudiées. Des systèmes continus approximant ces systèmes stables de commande à données discrètes sont construits et leurs propriétés transitoires sont discutées.

## Zusammenfassung

Der Aufsatz beschreibt eine neue Klasse von diskontinuierlich arbeitenden Regelungssystemen, in denen pulsfrequenzmodulierte Signale auftreten. Diese Darstellung ergab sich aus den Untersuchungen der Informationsübertragung in Nervenbahnen in physiologischen Regelungssystemen.

Ein Integral-Pulsfrequenz-Modulator wird definiert, der immer dann einen Impuls aussendet, wenn das Integral der Eingangsgröße einen Schwellwert erreicht hat. Die Frequenz der Ausgangsimpulsfolge ist linear zur Eingangsgröße. Modulatoren, die Impulsfolgen mit einem oder zwei Vorzeichen erzeugen, werden eingeführt. Die Ableitung der angenäherten Frequenzcharakteristiken zeigt, daß sie die Merkmale der diskontinuierlichen und diskreten Systeme haben und von der Eingangsgröße, der Frequenz und vom Anfangsphasenwinkel abhängen. Regelkreise, die in zwei parallelen Vorwärtszweigen Modulatoren mit einer Polarität besitzen, werden beschrieben und einige Stabilitätsbetrachtungen durchgeführt. Kontinuierliche Systeme, die als Approximation dieser stabilen diskontinuierlichen Systeme dienen, werden abgeleitet und ihre Übergangseigenschaften diskutiert.

## Introduction

It has been well observed through experiments that when a stimulation is exerted upon a sensory organ, some form of energy absorption and transduction will occur, thus a neural signal will be generated and transmitted along the nerve fibres linking it to some reflex centre where a decision is made and a control effort is issued to respond to the stimulation<sup>1</sup>. The neural signal in a single nerve fibre is a train of pulses of approximately constant magnitude and uniform shape, but its instantaneous frequency varies with the strength of the stimulation. Pulse frequency modulation is thus believed to exist in the neural communication networks of physiological control systems<sup>2</sup>. In the past, only a little work was done on pulse frequency modulation<sup>3</sup>. Integral pulse frequency modulation has been proposed by Li and Meyer<sup>2, 4, 5</sup>. It is derived from an idealized model of a relaxation oscillator which is assumed to take place in the transduction process<sup>2</sup>.

Physiological control systems contain communication channels that consist of hundreds and thousands of parallel nerve fibres, in each of which the signal is carried by pulse frequency modulation. These parallel channels serve to enlarge the dynamic range of the transmitted signal and improve the reliability. The control systems described in this paper contain only single 'fibres' and thus represent a drastic abstraction of the physiological case.

## Integral Pulse Frequency Modulation

In integral pulse frequency modulation, the modulator emits a pulse of fixed magnitude and width whenever the time integral of its activating signal reaches a threshold magnitude. The integrator is reset after each pulse so that the successive integration always starts from zero. An arbitrary input signal is thus converted into a pulse train whose repetition frequency varies directly with the magnitude of the activating signal. A simple linear and time-invariant low-pass filter is used as the demodulator whose output is a summation of pulse responses.

### Double-signed Integral Pulse Frequency Modulator

Let  $e_i(t)$  be the activating signal to the integral pulse frequency modulator (IPFM), and  $e_f(t)$  be its output. The output pulses are assumed to be of rectangular shape. The modulator output is defined by

$$e_f(t) = E_f \sum_{k=1}^{\infty} \operatorname{sgn}[\alpha(t_k)] [u(t - t_k) - u(t - t_k - \tau)] \quad (1)$$

or, if the pulse width  $\tau$  is very small, each pulse can be approximated by an equivalent impulse

$$e_f(t) = M \sum_{k=1}^{\infty} \operatorname{sgn}[\alpha(t_k)] \delta(t - t_k) \quad (2)$$

\* This study was supported in part by the grant B-2165 from the National Institutes of Health, U.S.A.

The time instant of the  $k$ th pulse or impulse,  $t_k$ , is determined from

$$|\alpha(t_k)| = \left| \int_{t_{k-1}}^{t_k} e_i(t) dt \right| = A \quad (k=1, 2, 3, \dots) \quad (3)$$

where  $u(t)$  = unit step function defined for  $t > 0$ ,  $E_f$  = pulse magnitude,  $\tau$  = pulse width,  $M = E_f \tau$  = strength of the pulse or the impulse,  $\delta(t)$  = unit impulse at  $t = 0$ , and

$$\begin{aligned} \text{sgn}[\alpha(t_k)] &= +1 & \text{if } \alpha(t_k) &= +A \\ &= -1 & \text{if } \alpha(t_k) &= -A \end{aligned}$$

Let  $T_k = t_k - t_{k-1}$  = time interval between the  $(k-1)$ th and the  $k$ th pulses,  $\tau < \min(T_1, T_2, \dots, T_k, \dots)$ . The instantaneous pulse frequency is defined as

$$f_{pk} = \frac{1}{T_k} \quad (4)$$

For a constant activating signal  $e_i(t) = E_i$ ,

$$\begin{aligned} |\alpha(t_k)| &= \left| \int_{t_{k-1}}^{t_k} E_i dt \right| = |E_i| (t_k - t_{k-1}) = |E_i| T_k = A \\ f_p &= f_{pk} = \frac{1}{T_k} = \frac{|E_i|}{A}, \text{ provided } f_p < \frac{1}{\tau} \text{ or } |E_i| < \frac{A}{\tau} \end{aligned} \quad (5)$$

Therefore, pulse frequency for a constant activating signal is a linear function of the magnitude of the activating signal with  $1/A$  as the proportionality constant, as shown in Figure 1(a).

Since pulses of both positive and negative signs can be generated depending upon the sign of the time integral of the activating signal, this type of modulator is called double-signed integral pulse frequency modulator (D-S IPFM).

#### Single-signed Integral Pulse Frequency Modulator

It is well known that the pulse signals in neural communication networks have only one sign. The single-signed integral pulse frequency modulator (S-S IPFM) is defined as one which can generate pulses of only one sign. This can be obtained by cascading a D-S IPFM to a linear rectifier as shown in Figure 1(b). Assume that the rectifier has a positive bias  $+E_b$ , a positive unity gain and a negative cut-off value  $-E_b$ . A linear rectifier with a negative unity gain and a positive cut-off value  $+E_b$  may also be used; this is introduced later when the reciprocally innervated parallel-path is dealt with. Let  $e(t)$  be the input signal to the S-S IPFM, the activating signal is

$$e_i(t) = \begin{cases} e(t) + E_b \geq 0 & \text{for } e(t) \geq -E_b \\ = 0 & \text{for } e(t) < -E_b \end{cases} \quad (6)$$

Hence, the output  $e_f(t)$  of the S-S IPFM is a pulse train of positive polarity only,

$$e_f(t) = E_f \sum_{k=1}^{\infty} [u(t - t_k) - u(t - t_k - \tau)] \quad (7)$$

or, by the impulse approximation,

$$e_f(t) = M \sum_{k=1}^{\infty} \delta(t - t_k) \quad (8)$$

where the pulse instant  $t_k$ , is determined from

$$\alpha(t_k) = \int_{t_{k-1}}^{t_k} e_i(t) dt = A \quad (9)$$

Let the input signal be a constant  $e(t) = E$ ,  $e_i(t) = E + E_b = E_i$

$$f_p = \frac{E + E_b}{A} = \frac{E}{A} + f_0 \text{ provided } 0 \leq E + E_b < \frac{A}{\tau} \quad (10)$$

where  $f_0$  is a fixed average pulse frequency due to  $E_b$

$$f_0 = \frac{E_b}{A} \quad (11)$$

The variation of the input signal  $e(t)$  about the bias level  $E_b$  is characterized by the variation of the instantaneous pulse frequency about  $f_0$ . [This principle has been applied successfully, through an obviously independent work, in an analogue tape recorder (series 100 and 200, Mnemotron, Inc., U.S.A.).]

Although the pulse frequency is a linear function of the activating signal magnitude, the integral pulse frequency modulator is a non-linear and discrete device because the output signal and the input signal do not satisfy the additive or homogeneous law. The precise analysis of the integral pulse frequency modulation is extremely difficult.

Consider a first order system  $a/(s+a)$  as a simple demodulator. Under a constant activating signal  $E_i$  to the modulator, it can easily be shown that the d.c. component of the output is equal to the activating signal  $E_i$  multiplied by a gain  $M/A$ , and the fundamental ripple of angular frequency  $\omega_r = 2\pi|E_i|/A$  is below 3 per cent of the d.c. level when  $10aA < |E_i| < A/\tau$ . Under this static condition, the IPFM may be approximately equivalent to a continuous element of the gain  $M/A$  for  $10 \cdot a < f_p < 1/\tau$ , where  $1/a$  may also be considered as the largest time constant in the demodulator. For S-S IPFM under arbitrary input  $e(t) \geq -E_b$ , the output of its demodulator contains a d.c. component corresponding to  $(M/A)E_b$ , the transmitted signal will be received by subtracting this d.c. level from the demodulator output.

Among the advantages of using the integral pulse frequency modulator are its simple mechanisms for modulation and demodulation, without the need of any standard clock frequency, and its good immunity to channel noise<sup>5</sup>.

#### Approximate Frequency Response of Single-signed Integral Pulse Frequency Modulator

Let  $e(t) = E_s \sin(\omega t + \beta)$ , where  $E_s$  is the input signal amplitude,  $\omega$  is the angular frequency, and  $\beta$  is the initial phase angle. Several pulse distributions of  $e_f(t)$  are illustrated in Figures 1(c)–(g). They may be periodic, subharmonic, or aperiodic, depending upon whether  $[\int_0^{2\pi/\omega} e_i(t) dt]/A$  is equal to an integer  $q$ , a rational number  $Q/P$ , or an irrational number. The output pulse train of the S-S IPFM under sinusoidal excitation is periodic at the signal frequency only for a set of discrete frequencies such that  $[\int_0^{2\pi/\omega} e_i(t) dt]/A = q$ , where  $q$  represents the number of pulses per signal period,

$$\left. \begin{aligned} \frac{\omega}{f_0} &= \frac{2\pi}{q} \\ &= \frac{1}{q} \left\{ \pi + 2 \left[ \sin^{-1} \frac{1}{E_s/E_b} + \left\{ \left( \frac{E_s}{E_b} \right)^2 - 1 \right\}^{\frac{1}{2}} \right] \right\} \end{aligned} \right\} \begin{cases} E_s \leq 1 \\ \text{for } E_s > 1 \end{cases} \quad (12)$$

The periodic pulse distribution does not undergo any transient process, it reaches an equilibrium configuration right after the

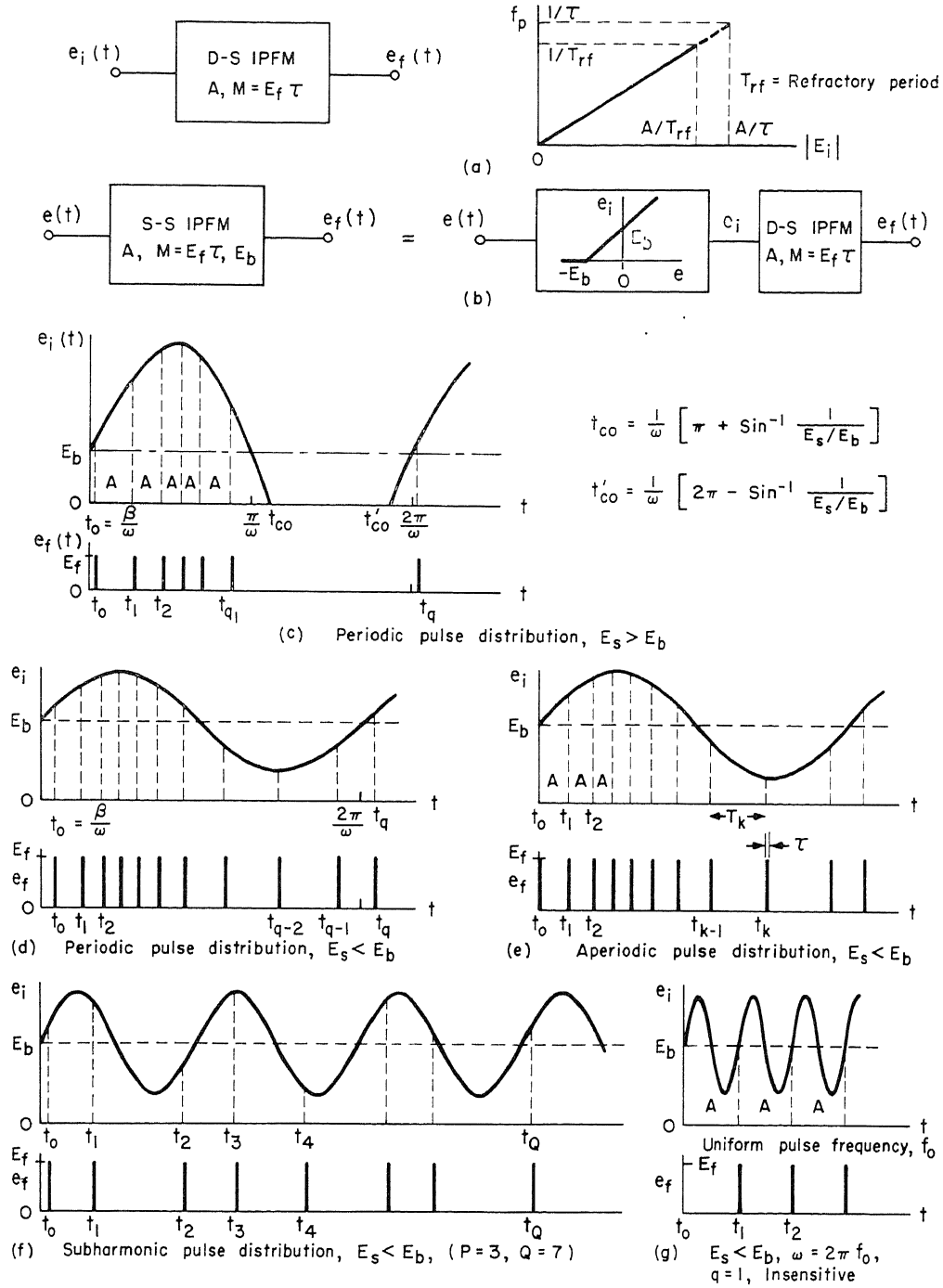


Figure 1. Integral pulse frequency modulators and pulse distributions of the S-S IPFM under a sinusoidal excitation  $e(t) = E_s \sin(\omega t + \beta)$

application of the input sinusoid. There is a critical value of  $\beta$ ,  $\beta_c$ , beyond which the periodic distribution will be equivalent to that for a certain  $\beta_{eq}$  where  $0 \leq \beta_{eq} < \beta_c$ .  $\beta_c$  is derived from

$$\int_{-\frac{\beta_c}{\omega}}^0 e_i(t) dt = \int_{-\frac{\beta_c}{\omega}}^0 [E_b + E_s \sin(\omega t + \beta_c)] dt = A, \quad (13)$$

$$\beta_c - \frac{E_s}{E_b} \cos \beta_c = \frac{\omega}{f_0} - \frac{E_s}{E_b}$$

There are infinitely many equilibrium configurations for  $\beta$  within the range of 0 to  $\beta_c$ . After any sudden disturbance to the input, the pulse instants within a signal period will settle immediately at their new relative locations corresponding to a different value of  $\beta$ . In this sense, each equilibrium periodic distribution is neutrally stable.

The effect of phase angle  $\beta$  of the input sinusoid  $e(t) = E_s \sin(\omega t + \beta)$  on the configuration of periodic pulse distribution may be exhibited by the time interval from a zero crossing

time  $t = 0$  to a reference time  $t_0 = (\beta/\omega)$  when it is assumed that  $e(t) = E_s \sin \omega t$ , as shown in *Figure 1*. For convenience, consider

$$e(t) = E_s \sin \omega t, \quad 0 \leq \omega t_0 = \beta < \beta_c \quad (14)$$

where  $\omega$  satisfies (12). Referring to *Figure 1(c)-(d)*, the pulse instants are determined from eqn (9),

for

$$\frac{E_s}{E_b} \leq 1, \quad \omega t_k - \frac{E_s}{E_b} \cos \omega t_k = k \frac{\omega}{f_0} + \beta - \frac{E_s}{E_b} \cos \beta, \quad (k=1, 2, \dots, q) \quad (15)$$

for

$$\left. \begin{aligned} \frac{E_s}{E_b} > 1, \quad \omega t_k - \frac{E_s}{E_b} \cos \omega t_k &= k \frac{\omega}{f_0} + \beta - \frac{E_s}{E_b} \cos \beta, \quad (k=1, 2, 3, \dots, q_1) \\ \omega t_{q_1+h} - \frac{E_s}{E_b} \cos \omega t_{q_1+h} &= (q_1+h) \frac{\omega}{f_0} + \pi - 2 \sin^{-1} \frac{1}{E_s/E_b} \\ &- 2 \left[ \left( \frac{E_s}{E_b} \right)^2 - 1 \right]^{\frac{1}{2}} + \beta - \frac{E_s}{E_b} \cos \beta, \quad (h=1, 2, \dots, q-q_1) \end{aligned} \right\} \quad (16)$$

where  $q_1$  is the number of pulses between  $t = t_0 + \tau = \beta/\omega + \tau$

$$\text{and} \quad t = t_0 + \tau = \frac{1}{\omega} \left[ \pi + \sin^{-1} \frac{1}{E_s/E_b} \right] + \tau$$

and is determined from  $q_1$  = the maximum positive integer equal to or less than

$$\frac{1}{\omega/f_0} \left[ \pi + \sin^{-1} \frac{1}{E_s/E_b} + \left[ \left( \frac{E_s}{E_b} \right)^2 - 1 \right]^{\frac{1}{2}} - \beta + \frac{E_s}{E_b} \cos \beta \right] \quad (17)$$

For very small  $\tau$ , the modulator output is represented by a periodic impulse train at the signal frequency,

$$e_f(t) = M \sum_{k=1}^q \delta(t - t_k) \text{ for } 0 \leq \frac{\beta}{\omega} < t \leq \frac{2\pi}{\omega} \leq \frac{2\pi + \beta}{\omega} \quad (18)$$

Its Fourier series is

$$e_f(t) = c_0 + \sum_{n=1}^{\infty} c_n \sin(n\omega t + \gamma_n) \quad (19)$$

where the coefficients are evaluated in the following way:

$$\begin{aligned} c_0 &= \frac{\omega}{2\pi} \int_{\frac{\beta}{\omega}}^{\frac{2\pi + \beta}{\omega}} e_f(t) dt = \frac{\omega}{2\pi} q M \\ c_n &= \frac{\omega}{\pi} M \left[ \left( \sum_{k=1}^q \cos n\omega t_k \right)^2 + \left( \sum_{k=1}^q \sin n\omega t_k \right)^2 \right]^{\frac{1}{2}} \quad (n=1, 2, 3, \dots) \\ \gamma_n &= \tan^{-1} \frac{\sum_{k=1}^q \cos n\omega t_k}{\sum_{k=1}^q \sin n\omega t_k} \quad (n=1, 2, 3, \dots) \end{aligned} \quad (20)$$

When the fundamental component is of paramount importance while the higher harmonics are negligible, as may be the case when the low-pass demodulator is used, the approximate frequency response,  $N(\omega/f_0, E_s/E_b, \beta)$ , or the describing function, of the S-S IPFM is defined as the complex gain of the modulator with regard to the fundamental component of  $e_f(t)$ ,

$$N\left(\frac{\omega}{f_0}, \frac{E_s}{E_b}, \beta\right) = \frac{c_1}{E_s} e^{j\gamma_1} \quad (21)$$

The normalized approximate frequency response or normalized describing function  $N_u(\omega/f_0, E_s/E_b, \beta)$  is

$$N_u\left(\frac{\omega}{f_0}, \frac{E_s}{E_b}, \beta\right) = \frac{1}{M/A} N\left(\frac{\omega}{f_0}, \frac{E_s}{E_b}, \beta\right) \quad (22)$$

$$\left. \begin{aligned} |N_u| &= \text{Modulus of } N_u\left(\frac{\omega}{f_0}, \frac{E_s}{E_b}, \beta\right) = \frac{1}{M/A} \frac{c_1}{E_s} \\ &= \frac{1}{\pi} \frac{\omega}{f_0} \frac{1}{E_s/E_b} \left[ \left( \sum_{k=1}^q \cos \omega t_k \right)^2 + \left( \sum_{k=1}^q \sin \omega t_k \right)^2 \right]^{\frac{1}{2}} \\ |N_u| &= \text{Argument of } N_u\left(\frac{\omega}{f_0}, \frac{E_s}{E_b}, \beta\right) \\ &= \gamma_1 = \tan^{-1} \frac{\sum_{k=1}^q \cos \omega t_k}{\sum_{k=1}^q \sin \omega t_k} \end{aligned} \right\} \quad (23)$$

For communication applications, it is required that  $E_s/E_b \leq 1$ . The normalized approximate frequency response can be computed at discrete frequencies for  $E_s/E_b \leq 1$  and  $\beta = 0^\circ$ . When  $q$  becomes increasingly larger, the difference between two consecutive neighbouring frequencies in the discrete set is smaller; in this sense, the approximate frequency response of the S-S IPFM may be thought of as being a function of a continuous variable  $\omega$ . The effect of  $\beta$  is to make  $N_u$  non-unique, this effect is reduced as  $\beta_c$  is reduced when  $q$  increases.  $N_u$  may exhibit either positive or negative phase shift. It oscillates about unity magnitude and null phase shift, and tends toward  $1 \angle 0^\circ$  as  $q$  increases or  $\omega/f_0$  decreases. For example, for  $q \geq 25$ ,  $\omega/f_0 \leq 0.08 \pi$ ,  $N_u$  can be approximated by  $1 \angle 0^\circ$  with an error bounded by 1 per cent in magnitude and about  $1.5^\circ$  in phase shift. Under this operating condition, the S-S IPFM, when followed by a low-pass demodulator, may be approximately equivalent to a continuous element of gain  $M/A$ . It is evident that if the parameter  $f_0$  is made larger it will improve the communication fidelity.

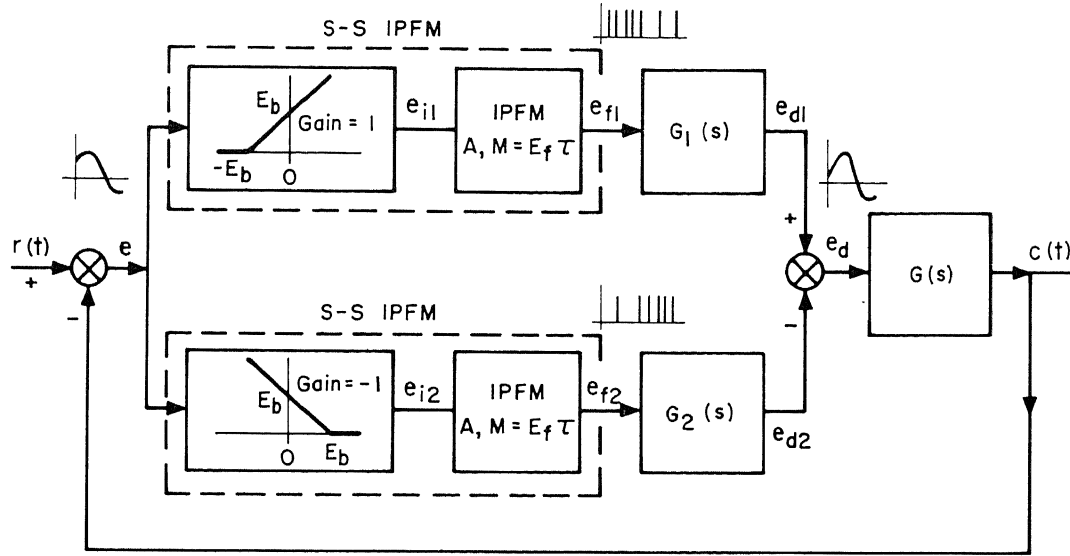
When  $q = 1$ , a uniform pulse train is obtained. In particular, if  $E_s/E_b \leq 1$ , the pulses are emitted at those instants where normal pulses would be even if  $e(t) = 0$ ; the S-S IPFM becomes insensitive.

### Parallel-Path Single-signed Integral Pulse Frequency Modulated Control Systems

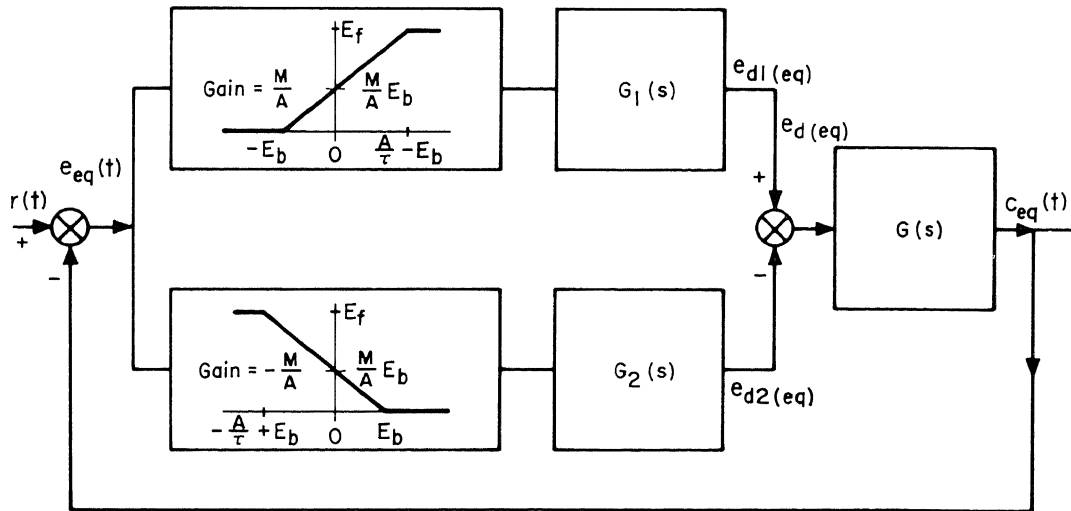
A new class of discrete-data feedback control systems is formed when integral pulse frequency modulators are introduced for system error modulation<sup>5</sup>. The parallel-path single-signed integral pulse frequency modulated control systems (P-P S-S IPFMCS) are presented here. The block diagram is shown in *Figure 2(a)*. There are two feedforward paths in parallel, each containing one S-S IPFM. These two modulators have the same values of  $E_b$ ,  $A$ ,  $E_f$  and  $\tau$ , but one has a linear rectifier with negative unity gain and positive cut-off value  $+E_b$ . The output of the first modulator,  $e_{f1}(t)$ , is to drive the system

$G_1(s)$ , and the output of the second modulator,  $e_{f2}(t)$ , the system  $G_2(s)$ .  $G_1(s)$  and  $G_2(s)$  serve as demodulators, their respective outputs  $e_{d1}(t)$  and  $e_{d2}(t)$  are subtracted one from the other;  $e_d(t) = e_{d1}(t) - e_{d2}(t)$  is to drive the final controlled element  $G(s)$ .  $r(t)$  and  $c(t)$  are the input and output of the system respectively. When the system error  $e(t)$ , also the input to the modulators, is zero, there is a constant pulse frequency  $f_0$  in each path; the average forces or torques developed in both

paths are virtually cancelled out at the subtractor if  $G_1(s)$  and  $G_2(s)$  are identical. When  $e(t)$  is non-zero, it increases the pulse frequency in one path while decreasing it in the other; the deviations of pulse frequencies in the two paths cause the developed force in one path to be higher and that in the other lower; both are in the same direction. The net control effort results from the pushpull action of the two paths. This characteristic is termed reciprocal innervation in the physiological



(a) PARALLEL-PATH SINGLE-SIGNED INTEGRAL PULSE FREQUENCY MODULATED CONTROL SYSTEM



(b) APPROXIMATELY EQUIVALENT CONTINUOUS SYSTEM OF PARALLEL-PATH SINGLE-SIGNED INTEGRAL PULSE FREQUENCY MODULATED CONTROL SYSTEM

Figure 2. Block diagrams of the parallel-path single-signed integral pulse frequency modulated control system and its approximately equivalent continuous system



context. Thus, the S-S IPFM's can be used for error modulation and at the same time the object can be controlled in either direction. The system is insensitive to small, high-frequency disturbances and behaves without a dead zone for slowly varying signals; in effect, it is actively controlled at all times.

Some salient phenomena have been noted when the pulse frequency is relatively low. This case is interesting because the nerve impulse in physiological systems varies approximately from 0 to 1,000 pulses/sec. While the actual physiological control loops are complex far beyond our present knowledge, this discussion is confined to some simple abstract models, as shown in Figure 2(a). All components except S-S IPFM's are assumed to be continuous and linear.  $G_1(s)$  and  $G_2(s)$  may be (a) identical, or (b) may simply have the same number of poles, the same number of zeros, and the same static gain:

$$\left. \begin{aligned} G_1(s) &= k_1 \frac{s^m + b_{11}s^{m-1} + \dots + b_{1m}}{s^n + a_{11}s^{n-1} + \dots + a_{1n}} \\ G_2(s) &= k_2 \frac{s^m + b_{21}s^{m-1} + \dots + b_{2m}}{s^n + a_{21}s^{n-1} + \dots + a_{2n}} \\ k_1 \frac{b_{1m}}{a_{1n}} &= k_2 \frac{b_{2m}}{a_{2n}} \end{aligned} \right\} \quad (24)$$

The former is called the balanced parallel-path, and the latter the specially unbalanced parallel-path. Their dynamic characteristics and an approximate method of analysis are presented in the following.

### Stability Analysis

Consider  $e(t) = 0$ ,  $e_{f1}(t)$  and  $e_{f2}(t)$  are two uniform pulse trains with pulse frequency  $f_0$ . The d.c. levels in  $e_{d1}(t)$  and  $e_{d2}(t)$  are the same; the ripples are of the repetition frequency  $f_0$ .  $e_d(t)$  and  $c(t)$  will contain a small oscillation of repetition frequency  $f_0$ , unless  $e_{f1}$  and  $e_{f2}$  are synchronized for a balanced parallel-path system. Although  $c(t)$  will be fed back,  $e_{f1}(t)$  and  $e_{f2}(t)$  are unaffected, providing that the feedback signal is bounded by  $E_b$ . This can be deduced from Figure 1(g). Thus, an equilibrium state can be established. This is defined as the static equilibrium state for the P-P S-S IPFMCS. There are many such static equilibrium states, each is neutrally stable, with regard to the degree of asynchronization of the two uniform pulse trains. After a sudden disturbance to the system,  $c(t)$  may finally tend toward one of its static equilibrium states. This is termed the 'almost asymptotic stability' of the P-P S-S IPFMCS. It is formally defined as follows. If the system output is disturbed at  $t_0$  from  $c(t) = 0$  with  $c(t_0) \leq \eta < \infty$ , there exists a small non-negative number  $\sigma$  depending on  $\eta$  such that  $\lim_{t \rightarrow \infty} |c(t)| \leq \sigma$ ;  $\sigma$  is bounded by a prescribed value, sufficiently small, which represents the maxima of the tolerable ripple in  $c(t)$  due to the two asynchronous pulse trains of frequency  $f_0$ . This definition is in contrast with the asymptotic stability of the continuous system where  $\sigma \rightarrow 0$  only. The concept 'almost asymptotic stability' is necessary for the description of control systems of this class.

### Application of Normalized Describing Function $N_u$

The normalized describing function  $N_u(\omega/f_0, E_s/E_b, \beta)$  has been derived in eqns (22) and (23). A portion of  $-(1/N_u)$  is

plotted in the polar form as shown in Figure 3.  $q$  contributes the discreteness of the consistent pair of  $(\omega/f_0, E_s/E_b)$ , for each of which, the variation of  $\beta$  from 0 to  $\beta_c$  traces out a curve for  $-1/N_u$  which shows a sharp change or discontinuity at a certain value of  $\beta$ . Through extensive numerical computations, all these two-parameter curves  $-[1/N_u(\omega/f_0, E_s/E_b)]$  seem to be symmetric with respect to the real axis, and may cross over one another. For  $q = 1$ ,  $0 \leq E_s/E_b \leq 1$  and  $\omega/f_0 = 2\pi$ ,  $-(1/N_u)$  simply occupies a circular region of radius 0.5 and centred at the origin, circles with small radii correspond to small  $E_s/E_b$ ; for  $q = 1$ ,  $E_s/E_b > 1$  and  $\omega/f_0 > 2\pi$ , the mappings are concentric circular arcs with radii within the range of 0.5 to  $\pi/2$ . For  $q \geq 2$ , the curves are situated to the left of a boundary to which  $-[1/N_u(\omega/f_0 = \pi, E_s/E_b = 0)]$  is asymptotic. When  $E_s/E_b \leq 1$ , the curves are convergent toward the  $-1$  point as  $q$  increases or  $\omega/f_0$  decreases. When  $E_s/E_b > 1$  and for a given  $q$ , the curves corresponding to larger  $E_s/E_b$  and larger  $\omega/f_0$  are located further to the left and may cross the negative real axis to the left of the  $-1$  point. Because  $\tau > 0$ , pulses may theoretically reach complete saturation in the positive half cycle with no pulse in the negative half cycle of the sinusoid when  $E_s/E_b$  is large enough; the modulator characteristics would approach that of a relay of one polarity with small hysteresis. The limiting plot of  $-(1/N_u)$  is then a narrow region with negative real axis as its centre line and extending to the left of the  $-1$  point.

The above discussion can be similarly applied to the S-S IPFM whose linear rectifier has a negative unity gain; its normalized describing function has a negative sign to that given by eqn (22), in addition,  $\beta$  is changed to  $\beta + 180^\circ$  which will be made equivalent to a proper  $\beta_{eq} = \beta' < \beta_c$ .

The describing functions of the modulators in the parallel-path are respectively

$$N_1\left(\frac{\omega}{f_0}, \frac{E_s}{E_b}, \beta\right) = \frac{M}{A} N_u\left(\frac{\omega}{f_0}, \frac{E_s}{E_b}, \beta\right)$$

and

$$N_2\left(\frac{\omega}{f_0}, \frac{E_s}{E_b}, \beta\right) = -\frac{M}{A} N_u\left(\frac{\omega}{f_0}, \frac{E_s}{E_b}, \beta'\right)$$

For describing function analysis,  $R(j\omega) = 0$ ,  $C(j\omega) = -E(j\omega)$ ,

$$C(j\omega) = G(j\omega) [E_{d1}(j\omega) - E_{d2}(j\omega)]$$

$$\begin{aligned} &= \frac{M}{A} G(j\omega) \left[ G_1(j\omega) N_u\left(\frac{\omega}{f_0}, \frac{E_s}{E_b}, \beta\right) \right. \\ &\quad \left. + G_2(j\omega) N_u\left(\frac{\omega}{f_0}, \frac{E_s}{E_b}, \beta'\right) \right] E(j\omega) \end{aligned}$$

and

$$\begin{aligned} &\frac{M}{A} G(j\omega) \left[ G_1(j\omega) N_u\left(\frac{\omega}{f_0}, \frac{E_s}{E_b}, \beta\right) + G_2(j\omega) N_u\left(\frac{\omega}{f_0}, \frac{E_s}{E_b}, \beta'\right) \right] \\ &= -1 \end{aligned} \quad (25)$$

The signal fed back to the modulators is not purely sinusoidal,  $\beta$  and  $\beta'$  of the fundamental components in the two paths may change from cycle to cycle. For the parallel-path systems under study, the phasors of  $G_1(j\omega)$  and  $G_2(j\omega)$  are rather close for very large and very small  $\omega$ . It is proposed that let  $\beta = \beta'$  so the

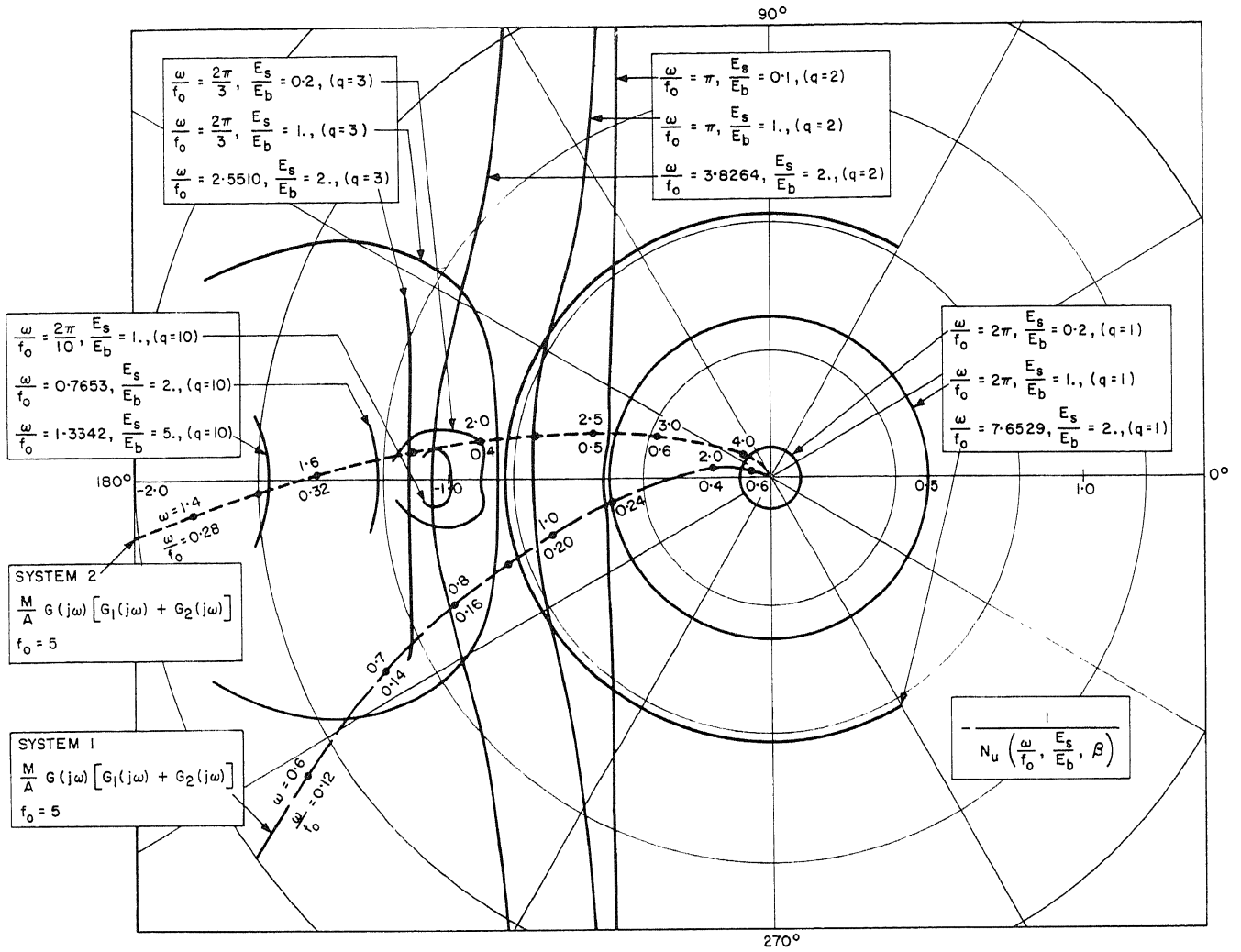


Figure 3. Describing function analyses of parallel-path single-signed integral pulse frequency modulated control systems

bracketed term in eqn (25) would be a maximum to cause the most severe stability situation. The critical equation is then

$$\frac{M}{A} G(j\omega) [G_1(j\omega) + G_2(j\omega)] = - \frac{1}{N_u \left( \frac{\omega}{f_0}, \frac{E_s}{E_b}, \beta \right)} \quad (26)$$

The transfer locus  $M/A G(j\omega) [G_1(j\omega) + G_2(j\omega)]$  is graded by  $\omega/f_0$  for a given  $f_0$  of the specific modulator. Any possible intersection of this transfer locus with any  $-1/N_u$  curve at a coincident value of  $\omega/f_0$  represents a possible oscillation at that angular frequency  $\omega$  and of the amplitude  $E_s$  determined from  $E_s/E_b$  graded on that intersecting curve of  $-1/N_u$ .

It is postulated that the approximated criterion for the 'almost asymptotic stability' of the balanced or specially unbalanced P-P S-S IPFMCS is that the transfer locus does not intersect with  $-1/N_u$  curves at any coincident frequency  $\omega/f_0$  except inside the circular region of a radius sufficiently smaller than 0.5 where  $q = 1$ , as illustrated by the system 1 in Figure 3. The latter intersection indicates the small steady state ripple, which may exist for such a stable system, of frequency  $f_0$  and

amplitude approximately bounded by the value of  $E_s$  determined from the intersecting circle;  $E_s$  is required to be below a specified value for smooth performance of the system.

If the transfer locus intersects, in addition, at other coincident frequencies  $\omega/f_0$  with  $-1/N_u$  curves for  $q \geq 2$ , as illustrated by the system 2 in Figure 3, the P-P S-S IPFMCS is considered unstable and will exhibit a strong self-sustained oscillation. There may be several intersecting points. Such an intersecting point, obtained only for a discrete set of  $(\omega/f_0, E_s/E_b)$ , would theoretically represent a pure sinusoidal oscillation. But realistically, the harmonics in the system output will be fed back to the modulators; not only the pulse distribution in the succeeding cycles will change, but the number of pulses per oscillation cycle will also vary. This effect is cumulative cycle after cycle. It may account for the mechanism of formation of almost sinusoidal, subharmonic, or aperiodic oscillation. The transfer locus of the system 2 crosses the negative real axis at about  $-1.46$  with  $\omega = 1.57$  rad/sec or  $\omega/f_0 = 0.314$ ; the intersections may occur in this neighbourhood. A self-sustained aperiodic oscillation is observed as shown in Figure 4(b) where  $\omega$  varies from 1.5 to 1.65 rad/sec and  $q$  varies from 26 to 20.

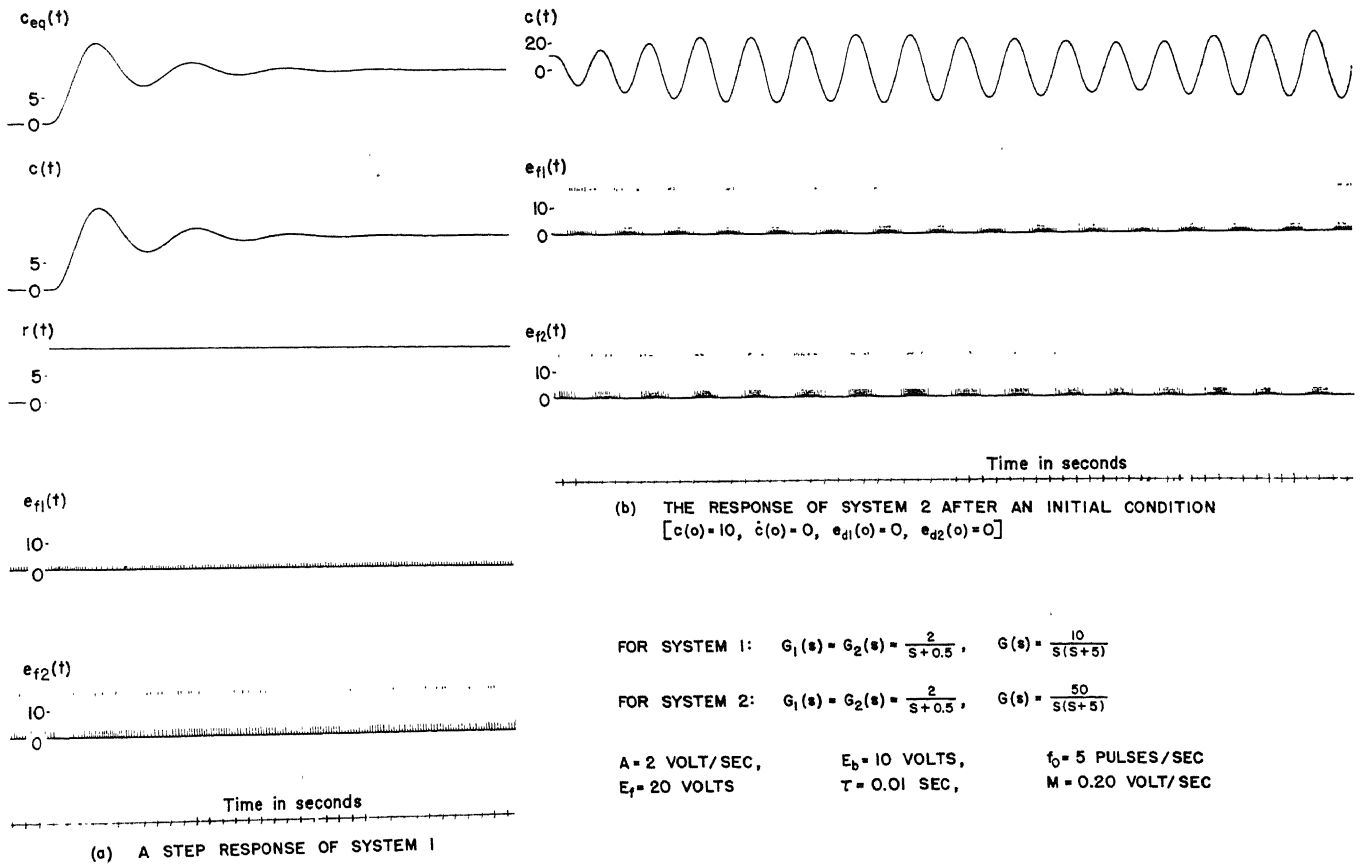


Figure 4. Time responses of two parallel-path single-signed integral pulse frequency modulated control systems. Their describing function analyses are shown in Figure 3

### Approximately Equivalent Continuous System

When a P-P S-S IPFMCS is 'almost asymptotically stable', an approximately equivalent continuous system can be postulated as shown in Figure 2(b) where a simple gain of  $M/A$  with saturation level  $E_f$  replaces each IPFM. The transient response due to the well defined deterministic input may be satisfactorily estimated from this continuous system approximation. An analogue computer study for the system 1 is illustrated in Figure 4(a),  $c(t)$  is shown almost the same as the output of the continuous system  $c_{eq}(t)$ .

### References

- JONES, R. W. Some properties of physiological regulators. *Automatic and Remote Control*, p. 675. 1961 London; Butterworths
- JONES, R. W., LI, C. C., MEYER, A. U. and PINTER, R. B. Pulse modulation in physiological systems, phenomenological aspects. *Trans. Inst. Radio Engrs BME-8* (1961) 59
- ROSS, A. E. Theoretical study of pulse-frequency modulation. *Proc. Inst. Radio Engrs* 37 (1949) 1277
- MEYER, A. U. Pulse frequency modulation and its effect in feedback systems. *Ph. D. Dissert.*, Northwestern University, U.S.A., 1961
- LI, C. C. Integral pulse frequency modulated control systems. *Ph. D. Dissert.*, Northwestern University, U.S.A., 1961

### DISCUSSION

E. I. JURY and T. PAVLIDIS, *University of California, Berkeley, California, U.S.A.*

The authors suggest a pulse frequency modulation scheme which is an idealized model of the neural communication. A more realistic model involves a more complicated form of pulse frequency modulation. One of the purposes of this discussion is to present such a generalized model. The form of this modulator is such that it emits a pulse when the convolution of the stimulus  $S(t)$  with the weighting function  $h(t)$  reaches a certain threshold, i.e.

$$\int_0^t S(\xi) h(t-\xi) d\xi = A \quad (1)$$

It is of interest to note that the authors' model is obtained as a special case from (1) by letting  $h(t) = 1$ . Furthermore, we obtain a system with a strength-duration curve which admits a non-zero threshold. This curve is well known to physiologists.

The general modulation scheme not only represents a better model of a neural communication but it also has advantages when used in feedback control systems. This becomes evident by observing in Figure A, that  $H(s)$  could play the role of a compensator. Furthermore, such a system could improve the filtering properties of the feedback system more than the IPFM.

In the stability investigation of IPFM, the authors have used the method of the describing function. This method while valid, leads to

a very complicated graphical and analytical procedure. It also fails to apply to the general case proposed by the discussors. Hence, a new stability procedure is attempted and has been successfully applied, which is based on the quasi-describing function. This function is derived by using a square-wave input to the non-linearity and obtaining an estimate of the ratio of the fundamental harmonic of the output

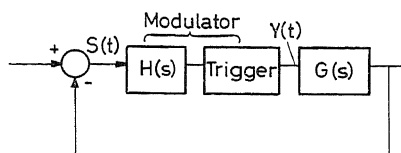


Figure A

of the non-linearity to the fundamental of the input. This is presented by the following equation:

$$Q(S, \omega) = \frac{2}{S_0 T} \int_0^T Y(t) dt \quad (2)$$

where  $\omega = 2\pi/T$ ,  $T$  is the square-wave period and  $S_0$  its amplitude. This general method has been applied to the authors' system and the results are shown in Table 1 in comparison with the results obtained by using the describing function and the exact experimental model. Although the describing function yields slightly more accurate results, the labour involved is considerably more than when using the quasi-describing function.

Table 1

Linear plant $G(s)$		Analogue computer	Describing function	Quasi-describ- ing function
5	$\omega$	0.965–0.998	0.93–0.97	0.95–1.0
$S(s + 0.2)$	$S$	3.0–3.2	2.7–2.96	$\sim 4$
10	$\omega$	1.3	1.2–1.3	1.0–1.4
$(s + 0.2)(s + 0.5)$	$S$	1.5–2.0	$\sim 1.95$	1.2–2.1
10	$\omega$	1.43–1.46	1.38–1.41	$\sim 1.4$
$s^2 + 0.24s + 0.16$	$S$	5.4–5.7	4.7–5	6.3

In conclusion, we commend the authors for their initial study of the system which stimulated our investigation of more complicated and realistic models. It is hoped that future studies will shed more light on the problem and also lead to further study of neural nets. Such a study will point out the value of this idealization.

## References

- <sup>1</sup> PAVLIDIS, T. Neural Pulse Frequency Modulated Control Systems, Int. Tech. Memo. M. 15, ERL, University of California, Berkeley, U.S.A. (April 1963)
- <sup>2</sup> BERTAUX, D. Mathematical model of a synapse, *M.S. Res. Proj.* Dept. Elec. Engng, University of California, Berkeley, U.S.A. (August 1963)

## C. C. LI, in reply

We thank Professor Jury and Mr. Pavlidis for their interesting discussion. Through their courtesy, we have received and read the complete report cited in their *Reference 1*. It is indeed encouraging to learn that our work has stimulated their research interests for generalization and improvement.

We also take this opportunity to mention two recent contributions related to our paper.

## References

- <sup>1</sup> MACKAY, D. M. and JEFFREYS, D. A. Continuous averaging system using magnetic tape. *Proc. 3rd Int. Conf. Medical Electronics, Lond.* (1960) 99
- <sup>2</sup> WEISS, R. F. A theory and simulation of rhythmic behaviour due to reciprocal inhibition in small nerve nets. *Proc. Spring Joint Computer Conf.* 1962. San Francisco, California (May 1962) 171

G. VOSSIUS, *Institut für animalische Physiologie, Ludwig-Rehn-Str. 14, Frankfurt/M., W. Germany*

Professor Li's paper is a theoretical treatment of the transfer behaviour of neural systems by considering the pulse frequency modulation of the information which occurs in such systems. In his theoretical abstraction, rectangular pulses are used for modulation; the line is ready to transmit a new pulse immediately after the end of the previous pulse. The nerve pulses have a quite different shape: a triangular pulse with a duration of about 1 msec is followed by a much longer refractory period during which the line will not transmit pulses. Thus, if the exciting frequency increases, only every second, third, etc. pulse will be followed by a nerve pulse. Furthermore, it is known that the shape of the neural pulse with its refractory period determines substantially the transfer properties of the neuron. Thus it is doubtful whether the theoretical considerations of Professor Li are transferable to the study of the behaviour of neural systems in this form. Possibly, thorough reconsideration will have to be given to the theory with regard to the physiological conditions.

## C. C. LI, in reply

The authors agree with Dr. Vossius' important remark from the realistic physiological point of view. As pointed out in the outset, the present paper concerns only the study of certain dynamic behaviours of the abstract systems including the proposed pulse-frequency modulators. The neural pulse in reality is neither a rectangular pulse nor a triangular one. Probably it could be described by the output from a non-linear network coupled to an IPFM such that the refractory period might be adequately taken care of. No doubt, extensive studies will have to be made along this direction. (Professor Jury mentioned that he has obtained some progress in this respect.)

## M. HAMZA, ETH Zurich, Switzerland

In the field of sampled-data control systems pulse width and pulse-frequency modulation are gaining in importance. This is due to their wide range of applications. Unfortunately, the analysis of systems employing such samplers is very difficult because the latter are highly non-linear. Very few publications have appeared on sampled-data control systems having a variable sampling frequency<sup>1</sup>. This work is novel and contains very useful ideas for which the authors are to be highly congratulated.

In the Institute of Automatic Control and Industrial Electronics of the ETH, control systems of the above type have been studied. I have considered pulse-width and pulse-frequency modulation as a means of compensating control systems<sup>2</sup>, and the results obtained were very encouraging, some of which—in addition to newer ones—will be published soon.

I would like to know if the authors have obtained any results on means of improving system response using the samplers they have described? Have they considered using the PPSSIPF modulator in an adaptive control system? Can they suggest any method for obtaining the time domain system response other than by analogue computer simulation?

Assuming the system could be approximated by a continuous one, what methods do they recommend for obtaining the system response? Finally, how should one select  $G_1(s)$ ,  $G_2(s)$ ?

## References

- <sup>1</sup> JURY, E. I. *Regelungstechnik* (March, 1963)
- <sup>2</sup> HAMZA, M. H. *Nonlinear, Signal Adaptive and Posicast Adaptive Control Systems*. Doctoral Dissert. 1963 Zurich; E.T.H.

C. C. LI, *in reply*

Dr. Hamza's discussion is greatly appreciated. In answer to the questions raised, I would say the following. The work reported in this paper essentially concerns the stability properties and salient features pertaining to the particular class of pulse-frequency modulated control systems. The IPFM itself is used as a device of transmitting signals,

not as a sampler. Studies have been made for a certain adaptive P-P-S-S IPFMCS incorporated with a reference model and for a varying rate sampled-data control system where an IPFM is used to trigger the sampler. With regard to the time domain system response, no method has yet been found to obtain the exact time response except the estimation based upon continuous system approximation, whenever permissible. In the system,  $G_1(s)$  and  $G_2(s)$  represent two continuous components satisfying conditions stated in the paper, their selection has not been discussed.

I am very glad to know that Dr. HAMZA has carried out researches in this field, and look forward to interchanging our new findings and ideas.

# Combination of Finite Settling Time and Minimum Integral of Squared Error in Digital Control Systems

V. PETERKA

## Summary

The design of the digital controller frequently follows the aim to obtain the system with the fastest response. By the practical application of this criterion, cases are often encountered where the transient error, caused by an input of typical form, has a very short duration but an inadmissible magnitude. This shortcoming is removed by the numerical method of controller design presented in the paper.

In this method the finite settling time is combined with the minimum of integral squared error. In this integral either the same importance is allotted to all errors during the control process, or only the errors occurring after the first sampling period are being considered. The physical meaning of the second case is the requirement of the computer liquidating the error, as far as possible, during one sampling period, and not instantaneously. The paper shows that good results can be obtained particularly by the second method. The method of calculating is arranged in such a way that both alternatives can be investigated simultaneously. The author succeeded in arranging the calculating procedure into a simple scheme. A numerical example is given.

## Sommaire

Les circuits numériques de réglage sont construits habituellement de manière, à ce que la durée du régime transitoire soit minimale. Dans l'application pratique de ce critère on trouve souvent des exemples où la durée du régime transitoire est courte, mais l'écart transitoire atteint des valeurs inadmissibles. Cet inconvénient est supprimé par la méthode numérique de calcul du régulateur décrite dans ce rapport.

Dans la méthode décrite on combine le critère de durée finie du régime transitoire avec la condition consistant à minimiser l'intégrale du carré de l'écart. Dans cette intégrale on tient compte soit de tous les écarts depuis le commencement du régime transitoire, soit seulement des écarts après le premier intervalle d'échantillonnage. On montre dans ce rapport que souvent cette seconde alternative conduit à de meilleurs résultats. On a réussi à arranger le calcul dans un schéma simple, permettant de calculer en même temps les deux cas. La méthode est démontrée sur un exemple numérique.

## Zusammenfassung

Der Einsatz von digitalen Reglern ist häufig damit begründet, Systeme mit minimaler Ausregelzeit zu bekommen. Bei der praktischen Anwendung dieses Kriteriums kommt es oft vor, daß für typische Eingangsgrößen zwar die Ausregelzeit sehr kurz ist, aber die Regelabweichung während des Übergangsvorganges unzulässig große Werte annimmt. Diese Arbeit enthält ein numerisches Verfahren, das diesen Nachteil beseitigt.

Die vorliegende Methode kombiniert die endliche Ausregelzeit mit der Forderung des minimalen Integralwertes der quadratischen Abweichung. Dieses Integral berücksichtigt entweder alle Fehler gleichwertig während des Regelvorganges oder nur diejenigen, die nach der ersten Tastperiode auftreten. Die physikalische Bedeutung des zweiten Falles ist die Forderung, daß der Digitalregler die Regelabweichung nicht sofort, sondern nach einer Tastperiode bestmöglich beseitigen soll. Der Aufsatz zeigt, daß sich gute Ergebnisse besonders

durch die zweite Methode erzielen lassen. Der Aufbau des Berechnungsverfahrens gestattet beide Möglichkeiten gleichzeitig zu berechnen. Ein numerisches Beispiel ist angeführt.

## Introduction

The aim frequently followed in the design of digital control systems is to eliminate the system error, caused by an input signal of a typical form (step, ramp, constant acceleration), within a minimum time<sup>1-3</sup>. Cases are often encountered in such systems with the fastest response where the transient error has a very short duration, but an inadmissible magnitude. This shortcoming can be removed by extending the response by one, two, or more sampling periods, as required, and by the application of a further criterion suppressing the system errors<sup>4</sup>. This article deals with a simple numerical method of digital controller design where the criterion of finite settling time has been combined with the minimum integral of squared error.

## The Statement of the Problem

Consider a control system compensated by a digital controller according to *Figure 1*. It is assumed that the transfer function of the plant is a rational fraction

$$S(p) = \frac{\sum_{v=0}^m b_v p^v}{\sum_{v=0}^n a_v p^v} = \frac{b(p)}{a(p)} = \frac{b(p)}{a_n \prod_{v=1}^n (p - p_v)} \quad (1)$$

with all its poles  $p_v$  in the left-hand side semi-plane  $p$ , or with maximum one pole equalling zero. For simplicity, let the problem be confined to an input signal having the form of a unit step

$$W(z) = \frac{1}{1 - z^{-1}}$$

and the holding device being of zero order

$$H(p) = \frac{1 - e^{-Tp}}{p}$$

Cases with another type of input signal, and with a holding device of a higher order, can be investigated in a similar way. It will also be assumed that the time required for the computing operation can be neglected, and the system has no dead time; however, it is possible to show that the consideration of both these lags is possible without any fundamental difficulties.

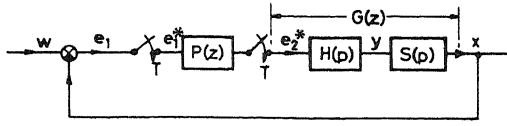


Figure 1

Let the pulse-transfer function of the continuously acting member be denoted

$$G(z) = \frac{B_0 + B_1 z^{-1} + \dots + B_n z^{-n}}{A_0 + A_1 z^{-1} + \dots + A_n z^{-n}} = \frac{B(z)}{A(z)} \quad (2)$$

The conditions of a finite settling time have been discussed in detail<sup>1-4</sup> and here they are stated only briefly in a form suited for the case.

For attaining a zero steady-state error at the sampling instants, after a finite number of sampling periods and under the conditions stated above, it is necessary that the overall pulse-transfer function

$$F(z) = \frac{P(z)G(z)}{1 + P(z)G(z)} \quad (3)$$

should be a polynomial in  $z^{-1}$ , and

$$F(1) = 1 \quad (4)$$

If the intersampling ripples also are to be eliminated, it is necessary to attain the settling of the manipulated variable  $y(t)$ . This will happen, if the pulse-transfer function  $E_2(z)/W(z)$  is also a polynomial in  $z^{-1}$ . The equation for this transfer function can be modified by the relations

$$E_2(z) = \frac{X(z)}{G(z)}, \quad F(z) = \frac{X(z)}{W(z)}$$

into the form

$$\frac{E_2(z)}{W(z)} = \frac{F(z)}{G(z)} = \frac{F(z)A(z)}{B(z)} \quad (5)$$

It follows from eqn (5) that all conditions stated will be fulfilled, if the overall pulse-transfer function has the form

$$F(z) = \frac{1}{B(1)} B(z) D(z) \quad (6)$$

where

$$D(z) = D_0 + D_1 z^{-1} + \dots + D_L z^{-L} \quad (7)$$

is a selectable polynomial for which

$$D(1) = \sum_{i=0}^L D_i = 1 \quad (8)$$

From relations (3) and (6) the necessary pulse-transfer function of the digital computer follows

$$P(z) = \frac{D(z)A(z)}{B(1) - D(z)B(z)} \quad (9)$$

If  $D(z) = 1$  is selected the system will have the fastest response, nevertheless the transient error can reach an inadmissible

magnitude as shown in the example that follows. Therefore let polynomial  $D(z)$  be of the general order of  $L$ , and state the problem as the determination of the coefficients  $D_0, D_1, \dots, D_L$  of the polynomial with the integral of the squared error

$$J = \int_0^\infty q(t) e_1^2(t) dt \quad (10)$$

having a minimum value.

Now it remains to select a suitable weighting function  $q(t)$ , so the following cases will be investigated.

In the first the same importance is allotted to all errors during the control process and  $q(t) = 1$  is selected, Figure 2(a). However, this selection need not be necessarily the most advantageous, namely the largest share in integral (10) belongs to errors at the beginning of the control process that cannot be physically eliminated in plants with a step function response starting from the origin. The minimalization of integral (10) can produce rather large overshoots that are not always desirable.

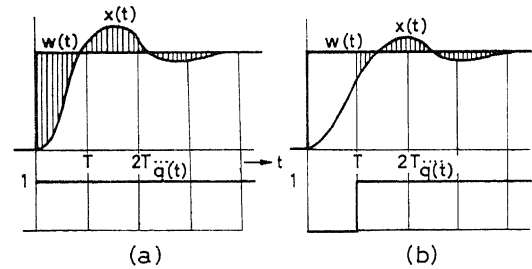


Figure 2

For this reason it is necessary to investigate the second case where no errors in the first sampling period are contained in integral (10), i.e. the weighting function is selected in the form of a unit-step function in time  $T$ ,  $q(t) = 1 (t - T)$  Figure 2(b). The physical meaning of this condition is the requirement of the computer liquidating the error, as far as possible, during one step, and not instantaneously as demanded in the former case. That is to say, in this second case the requirement put forward is less severe, and technically easier to realize.

The method of calculation is arranged in such a way that both cases can be investigated simultaneously, and thus it is possible to reach a decision in favour of the case that is more beneficial at the given concrete application.

### The Survey of Results

Coefficients  $D_1, D_2, \dots, D_L$  of the selectable polynomial (7), fulfilling the condition of the minimum integral (10) of the squared error, can be found by the solution of the system of linear equations

$$[K_{rs}][D_s] = -[R_{r0}] \quad (11)$$

where the square matrix  $[K_{rs}]$  is symmetrical the elements of which, and also the elements of column matrix  $[R_{r0}]$ , are independent of the selected degree  $L$  of polynomial  $D(z)$ . Two different cases are considered in the calculation of the elements of matrices  $[K_{rs}]$  and  $[R_{r0}]$ : (a) the transfer function  $S(p)$  of the plant has no zero pole, and (b)  $S(p)$  has one zero pole.

## Case (a)

In the case of the transfer function  $S(p)$  having all its poles different from zero, the step function response of the system is given by the equation

$$s(t) = \mathcal{L}^{-1} \left\{ \frac{S(p)}{p} \right\} = C_0 + \sum_{v=1}^n C_v e^{p_v t} \quad (12)$$

In this, and in all other equations that follow, the assumption is made that all the poles differ from each other. The case of multiple poles can be introduced by means of limits. The elements of matrices  $[K_{rs}]$  and  $[R_{r0}]$  are calculated by the following procedure. First of all the following expressions are solved numerically

$$\theta(k) = \sum_{v=1}^n \rho_v z_v^k, \quad k=0, 1, 2, \dots, n+L \quad (13)$$

where

$$z_v = e^{p_v T} \\ \rho_v = C_v \left( \frac{C_0}{p_v} + \delta_v \right), \quad \delta_v = \sum_{\mu=1}^n \frac{C_\mu}{p_v + p_\mu} \quad (14)$$

The calculation is made for  $k = 0, 1, 2, \dots, n+L$ , where  $L$  is the selected degree of polynomial  $D(z)$ .

The procedure is continued in such a way that all elements of the same row of matrix  $[K_{rs}]$ , and also of matrix  $[R_{r0}]$ , are calculated simultaneously for the weighting function  $q(t) = 1$ , and also for  $q(t) = 1(t-T)$ . As  $[K_{rs}]$  is a symmetrical matrix, it is sufficient to calculate the numerical values only of the elements lying below and on the main diagonal.

In order to calculate the elements of the  $r$ th row, the following equations have to be solved numerically

$$\begin{aligned} {}^r\Gamma_{-k} &= \theta(k) - \theta(r+k), \quad k=0, 1, 2, \dots, n \\ {}^r\Gamma_k &= c \cdot \min(k, r) + \theta(k) - \theta(|r-k|), \quad k=1, 2, \dots, n+r \end{aligned} \quad (15)$$

where

$$c = C_0^2 T \quad (16)$$

and  $\min(k, r)$  denotes the lower of the numbers  $k, r$ .

The figures  ${}^r\Gamma_{-n} \dots {}^r\Gamma_{n+r}$  obtained in this way are entered into column (a) shown in Table 1.

Table 1

(a)	(b)
${}^r\Gamma_{-n}$	${}^rU_0 \quad R_{r0}$
${}^r\Gamma_{-n+1}$	${}^rU_1 \quad R_{r1}$
$\vdots$	$\vdots$
${}^r\Gamma_{-n+k}$	${}^rU_s \rightarrow R_{rs}$
${}^r\Gamma_{-n+k+1}$	${}^rU_{s+1}$
$\vdots$	$\vdots$
${}^r\Gamma_{k-1} \quad {}^rU_{k-1}$	$\vdots$
${}^r\Gamma_k \rightarrow {}^rU_k$	${}^rU_{s+n-1}$
$\vdots$	${}^rU_{s+n}$
${}^r\Gamma_{n+r}$	$\vdots$
	${}^rU_{n+r}$

A slip of paper is laid beside the column with the coefficients  $A_n, A_{n-1}, \dots, A_0$  of the denominator of pulse-transfer function  $G(z)$  written on it one below the other. The product of figures lying beside each other (see column (a) of Table 1) supplies the value of

$${}^rU_k = \sum_{i=0}^n A_i {}^r\Gamma_{k-i} \quad (17)$$

which is then entered in the next column into the row containing the coefficient  $A_0$ . All the required values of  ${}^rU_k$  ( $k = 1, 2, \dots, n+r$ ) are then obtained by a gradual shifting of the paper slip with the coefficients  $A_i$  written on it.

The next operation represented by

$$R_{rs} = \sum_{i=0}^n A_i {}^rU_{i+s} \quad (18)$$

is performed again by using the paper slip with coefficients  $A_i$ . However, this time the coefficients are written in the opposite order as can be seen from column (b) of Table 1 where the paper slip is drawn in a position at which the numerical value of  $R_{rs}$  is being determined. The value  $R_{r0}$  is already the requested element of column matrix  $[R_{r0}]$  for the weighting function  $q(t) = 1$ . The elements of matrix  $[K_{rs}]$  are obtained simply as the difference

$$K_{rs} = R_{rs} - R_{r0} \quad (19)$$

The correctness of the calculations made so far can be checked by the relation

$$K_{rr} + 2R_{r0} = rc \left( \sum_{i=0}^n A_i \right)^2 \quad (20)$$

The elements of matrices  $[\hat{K}_{rs}]$  and  $[\hat{R}_{r0}]$  pertaining to the weighting function  $q(t) = 1(t-T)$  are obtained by adding the same figure of  $\Delta K$  respectively  $\Delta R$  to all elements of matrices  $[K_{rs}]$  and  $[R_{r0}]$  respectively.

$$\begin{aligned} \hat{K}_{rs} &= K_{rs} + \Delta K \\ \hat{R}_{r0} &= R_{r0} + \Delta R \end{aligned} \quad (21)$$

The figures to be added are obtained from

$$\begin{aligned} \Delta K &= -\lambda A_0^2 \\ \Delta R &= \lambda A_0^2 + \kappa A_0 \sum_{i=1}^n A_i \end{aligned} \quad (22)$$

where

$$\begin{aligned} \lambda &= c - \theta(0) - C_0 \sum_{v=1}^n \frac{C_v}{p_v} + \sum_{v=1}^n C_v z_v \left( \frac{2C_0}{p_v} + \delta_v \right) \\ \delta_v &= \sum_{\mu=1}^n \frac{C_\mu z_\mu}{p_v + p_\mu} \\ \kappa &= \sum_{v=1}^n \frac{C_v}{p_v} (1 - z_v) - c \end{aligned} \quad (23)$$

## Case (b)

In the case where one pole of transfer function  $S(p)$  lies in the origin, the step function response of the system is given by the equation

$$s(t) = \mathcal{L}^{-1} \left\{ \frac{S(p)}{p} \right\} = C_1 t + C_0 + \sum_{v=1}^N C_v e^{p_v t}, \quad N = n-1 \quad (24)$$



The elements of matrices  $[K_{rs}]$  and  $[R_{r0}]$  are calculated by the same method, only some values are calculated according to changed formulae.

Now, the numerical values of  $\theta(k)$  for  $k = 1, 2, \dots, N+L$  (where  $N = n-1$  is the number of non-zero poles) are obtained from the relations

$$k > 0, \quad \theta(k) = \sum_{v=1}^N \rho_v z_v^{k-1} \quad (25)$$

where

$$\rho_v = C_v (1 - z_v)^2 \left( \frac{C_{-1}}{p_v^2} - \frac{C_0}{p_v} - \delta_v \right) \quad (26)$$

$$z_v = e^{p_v T}, \quad \delta_v = \sum_{\mu=1}^N \frac{C_\mu}{p_v + p_\mu}$$

The value of  $\theta(0)$  is calculated separately from

$$\theta(0) = \frac{C_{-1}^2 T^3}{6} - C_0^2 T - 2 C_{-1} T \sum_{v=1}^N \frac{C_v}{p_v} - 2 \sum_{v=1}^N \frac{\rho_v}{1 - z_v} \quad (27)$$

The further procedure of calculation remains the same, except that for  $c$  we substitute everywhere

$$c = C_{-1}^2 T^3 \quad (28)$$

and instead of coefficients  $A_i$  ( $i = 0, 1, \dots, n$ ) we use everywhere the coefficients  $\bar{A}_i$  ( $i = 0, 1, 2, \dots, N$ ). Their relationship can be seen from the arrangement of the denominator of the pulse-transfer function  $G(z)$

$$A(z) = A_0 + A_1 z^{-1} + \dots + A_n z^{-n} \\ = (1 - z^{-1})(\bar{A}_0 + \bar{A}_1 z^{-1} + \dots + \bar{A}_N z^{-N}), \quad N = n-1 \quad (29)$$

This arrangement is made possible just because one pole of the transfer function  $S(p)$  equals zero.

The last difference in comparison with case (a) lies in the determination of the numerical values of  $\lambda$  and  $\kappa$  which are used in the determination of the matrices pertaining to the weighting function  $q(t) = 1$  ( $t = T$ ). They are calculated from the formulae

$$\lambda = \frac{c}{2} + C_0 C_{-1} T^2 - \theta(0) - \sum_{v=1}^N C_v (1 - z_v) \left( \frac{2 C_{-1} T}{p_v} - \delta_v \right) \\ \delta_v = \sum_{\mu=1}^N \frac{C_\mu (1 - z_\mu)}{p_v + p_\mu} \quad (30)$$

$$\kappa = \frac{c}{2} + C_0 C_{-1} T^2 - C_{-1} T \sum_{v=1}^N \frac{C_v}{p_v} (1 - z_v)$$

#### Example

In order to illustrate the method of calculation described generally in the preceding section, the calculation of a concrete case is given below. The transfer function of the plant is

$$S(p) = \frac{6p + 4.5}{(p+2)(p+1)(p+0.5)}$$

All poles of this transfer function are different from zero, the problem discussed is thus of the type of Case (a). The unit-step function response of the system is

$$s(t) = \mathcal{L}^{-1} \left\{ \frac{S(p)}{p} \right\} = C_0 + C_1 e^{p_1 t} + C_2 e^{p_2 t} + C_3 e^{p_3 t}$$

$$p_1 = -2; p_2 = -1; p_3 = -0.5; C_0 = 4.5; C_1 = 2.5; C_2 = -3; C_3 = -4.$$

The continuously acting member of the system has a pulse-transfer function

$$G(z) = \frac{B(z)}{A(z)} = \frac{1.309 z^{-1} - 0.092 z^{-2} + 0.248 z^{-3}}{1 - 1.110 z^{-1} + 0.355 z^{-2} - 0.030 z^{-3}}$$

Let the calculation of the coefficients of polynomial  $D(z)$  for its selected degree  $L = 1, 2, 3$  be presented. By solving eqns (14) and (13) we obtain

$$\delta_1 = \frac{79}{40}; \quad \delta_2 = \frac{10}{3}; \quad \delta_3 = 5$$

$$\rho_1 = -0.6875; \quad \rho_2 = 3.5; \quad \rho_3 = 16$$

k	0	1	2	3	4	5	6
$\theta(k)$	18.813	10.899	6.347	3.743	2.229	1.337	0.805

Table 2 contains the calculation of the elements of the second row ( $r = 2$ ) of matrices  $[K_{rs}]$  and  $[R_{r0}]$ .

Table 2

k	${}^2T_k$	${}^2U_k$	$R_{2k}$	$K_{2k}$
-3	2.406			
-2	4.118			
-1	7.156			
0	12.465	5.913	-0.694	
1	20.250	8.833	0.937	1.630
2	28.035	9.771	2.567	3.260
3	33.344	9.045		
4	36.382	8.720		
5	38.094	8.710		

The first column in Table 2 has been compiled according to eqns (15), the second and third have been calculated schematically according to Table 1. The fourth column containing elements  $K_{2k}$  has been obtained by means of relation (19).

The elements of the remaining two rows of matrices  $[K_{rs}]$  and  $[R_{r0}]$  are calculated in a similar way. As  $[K_{rs}]$  is a symmetrical matrix, it is sufficient to calculate only its elements lying to the left of the main diagonal and those on the diagonal itself. The correctness of the calculation is checked by substituting into relation (20) which is the means of checking almost all numerical operations represented in Table 2 including the compilation of the first column  ${}^rT_k$ .

By this method it has been possible to obtain a system of linear equations for the sought coefficients pertaining to the weighting function  $q(t) = 1$ :

$$\begin{bmatrix} 1.697 & 1.630 & 13.27 \\ 1.630 & 3.260 & 2.890 \\ 1.327 & 2.890 & 4.217 \end{bmatrix} \begin{bmatrix} D_1 \\ D_2 \\ D_3 \end{bmatrix} = \begin{bmatrix} 0.380 \\ 0.694 \\ 0.704 \end{bmatrix}$$

As the elements of matrices  $[K_{rs}]$  and  $[R_{r0}]$  are independent of the chosen degree  $L$  of polynomial  $D(z)$  the mere reduction of the respective matrices will suffice to meet the case of  $L = 1, 2$ .

By the solution of the above system of equations coefficients  $D_i$  are obtained for  $i \neq 0$ , while the coefficient  $D_0$  follows from condition (8)

$$D_0 = 1 - \sum_{i=1}^L D_i$$

In this way the following results have been obtained

$L$	$D_0$	$D_1$	$D_2$	$D_3$
3	0.758	0.046	0.140	0.057
2	0.768	0.038	0.194	
1	0.776	0.224		

In order to obtain the system of equations for the coefficients  $\hat{D}_i$  pertaining to the weighting function  $q(t) = 1(t - T)$  it will suffice, in accordance with relation (21), to add to each element of matrices  $[K_{rs}]$  and  $[R_{rs}]$  respectively the following quantities

$$\Delta K = -0.4548 \text{ and } \Delta R = -0.0646$$

obtained by the numerical solution of eqns (22) and (23). In this way one obtains

$L$	$\hat{D}_0$	$\hat{D}_1$	$\hat{D}_2$	$\hat{D}_3$
3	0.611	0.186	0.120	0.084
2	0.631	0.170	0.199	
1	0.642	0.358		

The pulse-transfer function of the continuously acting member of the system  $G(z) = B(z)/A(z)$  and the polynomial  $D(z)$ , the coefficients of which have just been calculated, determine completely the necessary transfer function (9) of the computer. The respective curves of the controlled variable  $x$  following the unit-step change of input signal  $w$  are represented in Figure 3 for the weighting function  $q(t) = 1$ , and in Figure 4 for the function  $q(t) = 1(t - T)$ .

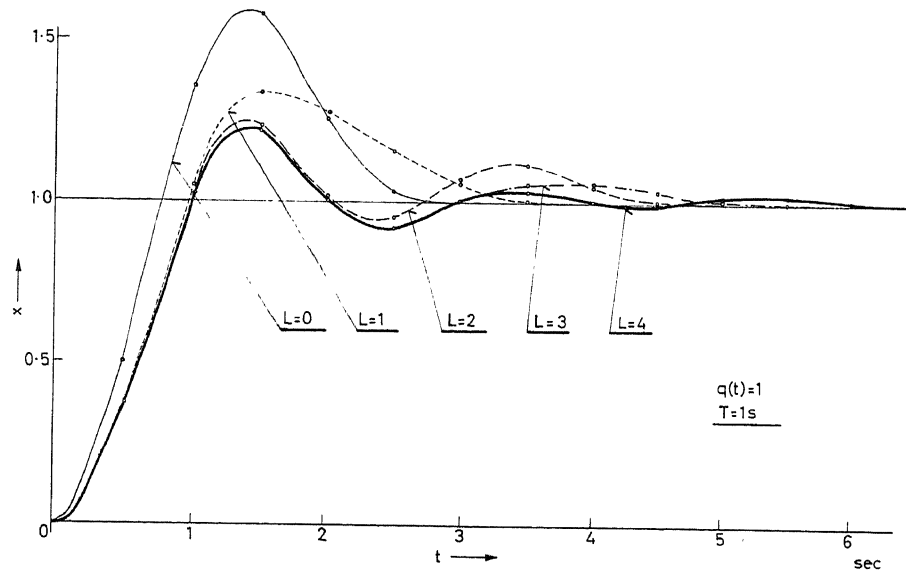


Figure 3

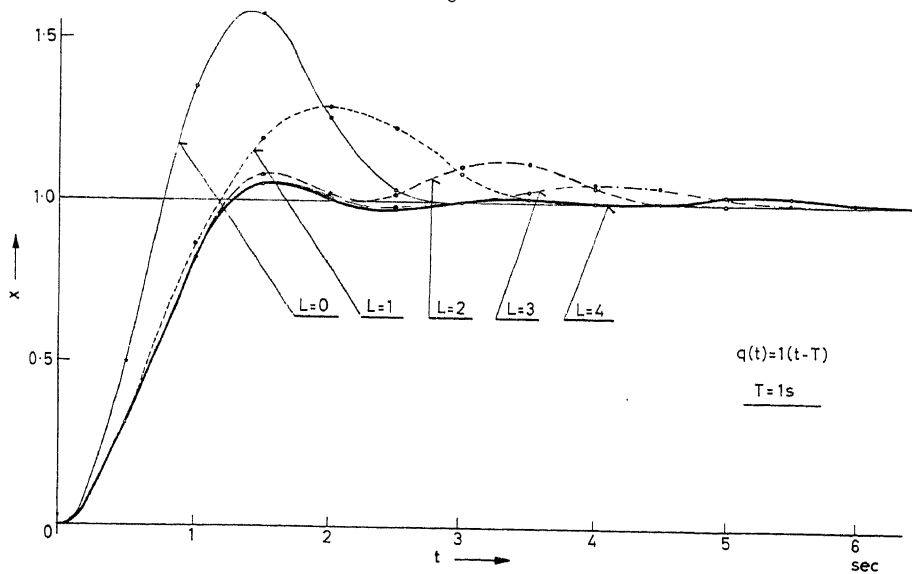


Figure 4

It can be seen from *Figures 3 and 4* that, compared with the minimum number of steps ( $L = 0$ ), a considerable improvement has been attained, especially in the case where in the minimalization of the integral of squared error the errors have been considered as occurring only after the first sampling period.

### Derivations and Proofs

First of all it will be proved that the above stated results hold for the case where all the poles of transfer function  $S(p)$  are different from zero.

The sequence of the increments of the variable  $e_2^*(t)$

$$\Delta e_2[i] = e_2^*(iT) - e_2^*(iT - T)$$

has, according to eqns (5) and (6), the  $z$ -transform of

$$\mathcal{L}\{\Delta e_2[i]\} = (1 - z^{-1}) E_2(z) = \frac{1}{B(1)} D(z) A(z) \quad (31)$$

From this  $z$ -transform it follows obviously

$$\Delta e_2[i] = \frac{1}{B(1)} \sum_{s=0}^L D_s A_{i-s} \quad (32)$$

where  $A_k = 0$  for  $k < 0$  and  $k > n$ ;  $\Delta e_2[i] = 0$  for  $i > n + L$ . Eqn (32) contains all the  $L + 1$  coefficients of polynomial  $D(z)$ ; however, only  $L$  of them can be selected, as it is necessary to fulfil condition (8) that is  $D(1) = 1$ . For the purpose of fulfilling this condition let coefficient  $D_0$  be detached

$$D_0 = 1 - \sum_{s=1}^L D_s \quad (33)$$

and eliminated from eqn (32)

$$\Delta e_2[i] = \frac{1}{B(1)} \left[ \sum_{s=1}^L D_s (A_{i-s} - A_i) + A_i \right] \quad (34)$$

Now, the time curve of the manipulated variable  $y(t)$  is dissolved into the sum of unit-step functions

$$y(t) = \sum_{i=0}^{n+L} \Delta e_2[i] 1(t - iT)$$

and the curve of the controlled variable  $x(t)$  can then be represented by the superposition of the unit-step responses

$$x(t) = \sum_{i=0}^{n+L} \Delta e_2[i] s(t - iT) \quad (35)$$

Then for error  $e_1(t)$  it holds that

$$\begin{aligned} e_1(t) &= 1 - x(t) = 1 - \sum_{i=0}^{n+L} \Delta e_2[i] s(t - iT) \\ &= \sum_{i=0}^{n+L} \Delta e_2[i] \bar{s}(t - iT) \end{aligned} \quad (36)$$

where

$$\bar{s}(t - iT) = s(\infty) - s(t - iT)$$

$$t > iT, \bar{s}(t - iT) = - \sum_{v=1}^n C_v e^{p_v(t - iT)}$$

$$t < iT, \bar{s}(t - iT) = C_0$$

If the integral of squared error (10) has a minimum value, the coefficients of polynomial  $D(z)$  must fulfil the equations

$$\frac{\partial J}{\partial D_r} = 0, \quad r = 1, 2, \dots, L \quad (38)$$

According to the above indicated derivative of the integral it follows

$$2 \int_0^\infty q(t) e_1(t) \frac{\partial e_1(t)}{\partial D_r} dt = 0 \quad (39)$$

The necessary partial derivative is determined from relations (36) and (34)

$$\frac{\partial e_1(t)}{\partial D_r} = \frac{1}{B(1)} \sum_{j=0}^{n+L} (A_{j-r} - A_j) \bar{s}(t - jT) \quad (40)$$

By substituting (40) and (36) together with (34) into condition (39), and by altering the sequence of addition, it follows

$$\begin{aligned} \int_0^\infty q(t) \left\{ \sum_{s=1}^L D_s \sum_{i=0}^{n+L} \sum_{j=0}^{n+L} (A_{i-s} - A_i) (A_{j-r} - A_j) \bar{s}(t - iT) \right. \\ \left. \bar{s}(t - jT) + \sum_{i=0}^{n+L} \sum_{j=0}^{n+L} A_i (A_{j-r} - A_j) \bar{s}(t - iT) \bar{s}(t - jT) \right\} dt = 0 \end{aligned} \quad (41)$$

Under the accepted pre-condition of transfer function  $S(p)$  having all its poles in the left semi-plane the integrals

$$\sigma_{ij} = \int_0^\infty q(t) \bar{s}(t - iT) \bar{s}(t - jT) dt \quad (42)$$

converge and in eqn (41) the integral of the sum can be expressed as the sum of the integrals

$$\begin{aligned} \sum_{s=1}^L D_s \sum_{i=0}^{n+L} \sum_{j=0}^{n+L} (A_{i-s} - A_i) (A_{j-r} - A_j) \sigma_{ij} \\ + \sum_{i=0}^n \sum_{j=0}^{n+L} A_i (A_{j-r} - A_j) \sigma_{ij} = 0 \end{aligned}$$

and in the abbreviated form

$$\sum_{s=1}^L D_s K_{rs} + R_{r0} = 0, \quad r = 1, 2, \dots, L \quad (43)$$

with the following denotations

$$K_{rs} = \sum_{i=0}^{n+L} \sum_{j=0}^{n+L} (A_{i-s} - A_i) (A_{j-r} - A_j) \sigma_{ij} \quad (44)$$

$$R_{r0} = \sum_{i=0}^n \sum_{j=0}^{n+L} A_i (A_{j-r} - A_j) \sigma_{ij} \quad (45)$$

For the degree  $L$  of the selectable coefficients  $D_s$  the system of linear equations is thus obtained that can be written in the matrix form (11) as

$$[K_{rs}] [D_s] = -[R_{r0}]$$

By interchanging the subscripts in relation (44) it can be easily proved that  $[K_{rs}]$  is a symmetrical matrix.

By the solution of integral (42) it follows for the case of weighting function  $q(t) = 1$

$$\sigma_{ij} = c \min(i, j) + C_0 \sum_{v=1}^n \frac{C_v}{p_v} - \theta(|j-i|) \quad (46)$$

where the function  $\theta(k)$  is determined by relation (13), and  $c$  according to relation (16). Integral (42) for the weighting function  $q(t) = 1 (t - T)$  is to be denoted by  $\hat{\sigma}_{ij}$ . It holds

$$\begin{aligned} i \neq 0, j \neq 0, \hat{\sigma}_{ij} &= \sigma_{ij} - c \\ \hat{\sigma}_{0k} &= \hat{\sigma}_{k0} = C_0 \sum_{v=1}^n \frac{C_v}{p_v} z_v - \theta(k) \end{aligned} \quad (47)$$

$$\hat{\sigma}_{00} = - \sum_{v=1}^n C_v z_v \delta_v$$

where  $\delta_v$  is determined by the second of relations (23), and  $z_v = e^{p_v T}$ .

The calculation of the elements of matrices  $[K_{rs}]$  and  $[R_{r0}]$  according to relations (44) and (45) would be very laborious. For this reason let some arrangements be introduced that will simplify this calculation considerably.

First, let us divide relation (44) into two terms

$$\begin{aligned} K_{rs} &= \sum_{i=0}^{n+L} \sum_{j=0}^{n+L} A_{i-s} (A_{j-r} - A_j) \\ &\quad - \sum_{i=0}^n \sum_{j=0}^{n+L} A_i (A_{j-r} - A_j) \sigma_{ij} \end{aligned}$$

Now, if the summing subscript  $i$  in the first member is shifted by  $s$ , i.e.  $i - s \rightarrow i$ , and considering that  $A_i = 0$  for  $i > n$  and  $i < 0$ , it follows

$$\begin{aligned} K_{rs} &= \sum_{i=0}^n \sum_{j=0}^{n+L} A_i (A_{j-r} - A_j) \sigma_{i+s, j} \\ &\quad - \sum_{i=0}^n \sum_{j=0}^{n+L} A_i (A_{j-r} - A_j) \sigma_{ij} \end{aligned}$$

According to (45) the second term equals  $R_{r0}$ , and let the first be denoted

$$R_{rs} = \sum_{i=0}^n A_i \sum_{j=0}^{n+L} (A_{j-r} - A_j) \sigma_{i+s, j} \quad (48)$$

In this way relation (19) has been obtained

$$K_{rs} = R_{rs} - R_{r0} \quad (49)$$

As the term  $R_{r0}$  represents a special case of  $R_{rs}$  with  $s = 0$ , it will suffice further to seek only the numerical solution of (48) for  $R_{rs}$ . Let it be written in the following form

$$R_{rs} = \sum_{i=0}^n A_i \sum_{j=0}^{n+L} A_j (\sigma_{i+s, j+r} - \sigma_{i+s, j})$$

If we denote

$${}^r U_{i+s} = \sum_{j=0}^{n+L} A_j (\sigma_{i+s, j+r} - \sigma_{i+s, j}) \quad (50)$$

we obtain relation (18)

$$R_{rs} = \sum_{i=0}^n A_i {}^r U_{i+s} \quad (51)$$

All values of  ${}^r U$  required for the calculation of the  $r$ th row of matrices  $[K_{rs}]$  and  $[R_{r0}]$  can be obtained as the product of the rectangular matrix

$$\boxed{\text{Eqn (52)}}^*$$

and of the column matrix

$$[A_i] = [A_0, A_1, \dots, A_n] \quad (53)$$

$$[{}^r U] = [\sigma] [A_i]$$

It will be proved that in the case of weighting function  $q(t) = 1$  matrix  $[\sigma]$  has all its elements lying on the lines parallel to the main diagonal of the same value. For the  $m$ th element of the  $k$ th parallel above main diagonal it holds

$$\begin{aligned} \sigma_{m, k+r+m} - \sigma_{m, k+m} &= cm + C_0 \sum_{v=1}^n \frac{C_v}{p_v} - \theta(k+r) \\ &\quad - \left[ cm + C_0 \sum_{v=1}^n \frac{C_v}{p_v} - \theta(k) \right] = \theta(k) - \theta(k+r) \end{aligned}$$

As this relation is independent of  $m$ , all elements lying on this parallel are equal, and they may be denoted by the same symbol

$${}^r \Gamma_k = \theta(k) - \theta(k+r)$$

Similarly it holds for the elements on the  $k$ th parallel below the main diagonal

$${}^r \Gamma_k = \sigma_{k+m, r+m} - \sigma_{k+m, m} = c \min(k, r) + \theta(k) - \theta(|r-k|)$$

For the main diagonal  $k = 0$ .

Due to this property of matrix  $[\sigma]$  it is possible to arrange the numerical solution of matrix product (51) into a scheme shown in Table 1 (a) which can be easily found by comparing both methods of calculation.

It remains yet to prove the validity of formulae (21), (22), and (23) by which the former results are to be corrected, if errors are being considered only after the first sampling period. By substituting into eqn (50) for  $\sigma_{ij}$  (46) the terms  $\hat{\sigma}_{ij}$  (47) calculated for the weighting function  $q(t) = 1 (t - T)$ , it can be seen that only the first column has been altered. Obviously it holds that

$${}^r \hat{U}_k = {}^r U_k + (\hat{\sigma}_{k, r} - \sigma_{k, r} - \hat{\sigma}_{k, 0} + \sigma_{k, 0}) A_0 \quad (54)$$

\* Eqn (52):

$$[\sigma] = \begin{bmatrix} \sigma_{0,r} & -\sigma_{0,0} & \sigma_{0,1+r} & -\sigma_{0,1} & \sigma_{0,2+r} & -\sigma_{0,2} & \dots & \sigma_{0,n+r} & -\sigma_{0,n} \\ \sigma_{1,r} & -\sigma_{1,0} & \sigma_{1,1+r} & -\sigma_{1,1} & \sigma_{1,2+r} & -\sigma_{1,2} & \dots & \sigma_{1,n+r} & -\sigma_{1,n} \\ \sigma_{2,r} & -\sigma_{2,0} & \sigma_{2,1+r} & -\sigma_{2,1} & \sigma_{2,2+r} & -\sigma_{2,2} & \dots & \sigma_{2,n+r} & -\sigma_{2,n} \\ \vdots & & \vdots & & \vdots & & & \vdots & \\ \sigma_{n+L,r} & -\sigma_{n+L,0} & \sigma_{n+L,1+r} & -\sigma_{n+L,1} & \sigma_{n+L,2+r} & -\sigma_{n+L,2} & \dots & \sigma_{n+L,n+r} & -\sigma_{n+L,n} \end{bmatrix} \quad (52)$$

When calculating the term in the parentheses, it is necessary to differentiate two cases:  $k > 0$  and  $k = 0$ . By substituting relations (46) and (47), we obtain in the first case the relation

$$k > 0, \hat{\sigma}_{k,r} - \sigma_{k,r} - \hat{\sigma}_{k,0} + \sigma_{k,0} = C_0 \sum_{v=1}^n \frac{C_v}{p_v} (1 - z_v) = \kappa$$

which is independent of  $k$ . Similarly for  $k = 0$

$$\begin{aligned} & \hat{\sigma}_{0,r} - \sigma_{0,r} - \hat{\sigma}_{0,0} + \sigma_{0,0} \\ &= -C_0 \sum_{v=1}^n \frac{C_v}{p_v} (1 - z_v) + \sum_{v=1}^n C_v z_v \delta_v + C_0 \sum_{v=1}^n \frac{C_v}{p_v} - \theta(0) = \kappa_0 \end{aligned}$$

With this notation the relation (54) may be rewritten in the form

$$\begin{aligned} k > 0, & \quad {}^r \hat{U}_k = {}^r U_k + \kappa A_0 \\ k = 0, & \quad {}^r \hat{U}_0 = {}^r U_0 + \kappa_0 A_0 \end{aligned} \quad (55)$$

For the verification of formulas (21) and (22) it will suffice to execute operations (51) and (49) with the relations (55), and to denote  $\kappa_0 - \kappa = \lambda$ .

The checking formula (20) can be verified by substituting relations (44) and (45) and by using the relation

$$\sigma_{i+r, j+r} - \sigma_{i, j} = r c$$

that follows from eqn (46).

In Case (b), with the transfer function  $S(p)$  having one zero pole, the continuously acting member of the system is astatic [ $s(\infty) = \infty$ ], and integrals (42) are not converging. It is possible to by-pass this difficulty, if the curve of the controlled variable is not represented as the superposition of unit-step responses, but as the superposition of responses to rectangular pulses. Otherwise the procedure of derivation is the same as in Case (a).

## References

- <sup>1</sup> RAGAZZINI, J. R. and FRANKLIN, G. F. *Sampled-data Control Systems*. 1958. New York; McGraw-Hill
- <sup>2</sup> JURY, E. I. *Sampled-data Control Systems*. 1958. New York; Wiley
- <sup>3</sup> TOUT, J. T. *Digital and Sampled-data Control Systems*. 1959. New York; McGraw-Hill
- <sup>4</sup> STREJC, V. Ensuring reliability in complex automation by automatic digital computers. *Automatizace*. V (1962) 5

# Oscillations Sous-harmoniques dans un Asservissement par Plus-ou-moins

J-C. GILLE, S. WĘGRZYN et J-G. PAQUET

## Summary

Conditions for the existence of  $n$ th-order subharmonics in an on-off control system whose linear part is of any order are established. In the case of a symmetrical relay without a dead zone use of Hamel locus of the system (instead of Tsytkin locus) immediately shows that only odd subharmonics can occur and provides a geometrical interpretation of commutation within half-period. It is thus seen that, for conventional servomechanisms, subharmonics, especially of high order, seldom appear, but the presence of backlash makes them more apt to appear; some particular types of systems easily generate subharmonics of all orders. Use of Hamel loci makes immediate generalization possible for the case in which a proportional-plus-derivative compensator is placed before the relay. Investigation of the global stability, or stability in the large, or the oscillations by direct simulation verifies these results. It shows that the border between two possible oscillations involves hysteresis, i.e., there is frequency entrainment; the part played by commutation within half-period appears as the cause of the disappearance of the higher-order subharmonic. Conditions for existence in the presence of asymmetry or of a dead zone are obtained as an extension of the Tsytkin method for determining forced oscillations from two families of loci. In the case of a non-symmetrical relay the problem is shown to be solved with one locus when the linear part of the servo system has integration or has a zero equal to zero.

## Sommaire

On donne des conditions nécessaires d'existence pour le sous-harmonique d'ordre  $n$  d'un asservissement par plus-ou-moins dont la partie linéaire est d'ordre quelconque. Dans le cas d'un relais symétrique sans seuil l'utilisation du lieu de Hamel (au lieu de celui de Cypkin) montre immédiatement que seuls peuvent exister les sous-harmoniques d'ordre impair; elle donne l'interprétation géométrique du phénomène de commutation prématurée. On voit ainsi que pour les asservissements usuels les sous-harmoniques, surtout d'ordre élevé, apparaissent difficilement, mais la présence d'hystérésis favorise leur production; certains types particuliers de systèmes présentent facilement des sous-harmoniques impairs de tous ordres. Grâce au lieu de Hamel ces résultats se généralisent sans calcul pour le cas où un correcteur dérivé précède le relais. L'étude de la stabilité globale des oscillations par simulation vérifie ces résultats: elle met en évidence une hystérésis dans la frontière entre deux oscillations possibles, i.e. l'existence d'un traînage de fréquence, et le rôle physique joué par la commutation prématurée dans la disparition de l'harmonique d'ordre plus élevé. Les conditions d'existence dans le cas dissymétrique ou avec seuil s'obtiennent par une extension de la méthode de Cypkin de détermination de l'oscillation forcée par deux familles de lieux. Dans le cas d'un relais dissymétrique avec une fonction linéaire possédant une intégration ou un zéro nul on montre qu'on peut résoudre le problème avec un lieu unique.

## Zusammenfassung

Es werden die notwendigen Bedingungen für die Existenz einer  $n$ -ten subharmonischen Schwingung eines Zweipunktregelsystems, dessen linearer Teil von beliebiger Ordnung ist, angegeben. Für ein Relais mit symmetrischer Kennlinie ohne Totzone (Ansprechempfindlichkeit) zeigt die Benutzung der Methode nach Hamel (nicht nach Zypkin)

sofort, daß nur subharmonische Schwingungen von ungerader Ordnung auftreten können; daraus bekommt man auch eine geometrische Deutung für das Umschalten während der ersten Hälfte der Periode. Es zeigt sich, daß bei üblichen Regelkreisen subharmonische Schwingungen, besonders höherer Ordnung, selten auftreten; Hysterese hingegen begünstigt deren Auftreten. In einigen besonderen Arten von Zweipunktregelsystemen treten leicht subharmonische Schwingungen von beliebiger ungerader Ordnung auf. Die Benutzung der Methode nach Hamel läßt diese Ergebnisse ohne Berechnung für den Fall verallgemeinern, daß ein PD-Kompensationsglied vor dem Relais liegt. Untersuchungen über die Stabilität im Großen mit Hilfe eines Analogrechners bestätigen diese Ergebnisse. Es zeigt sich, daß in dem Bereich zwischen zwei möglichen Schwingungen eine Hysterese Wirkung besteht, die eine Frequenzmitnahme erzeugt. Durch Umschalten während der ersten Hälfte der Periode verschwinden die subharmonischen Schwingungen höherer Ordnung. Bedingungen für das Auftreten subharmonischer Schwingungen bei einer asymmetrischen Relaiskennlinie oder bei einem Relais mit Ansprechempfindlichkeit erhält man durch Verallgemeinerung des Zypkinschen Verfahrens zur Bestimmung von erzwungenen Schwingungen durch zwei Ortskurvenscharen. Im Falle eines asymmetrischen Relais zeigt sich, daß das Problem mit Hilfe einer einzigen Ortskurve lösbar ist, wenn der lineare Teil des Regelsystems integrales Verhalten aufweist oder einen Pol im Nullpunkt besitzt.

## Introduction

L'existence d'oscillations sous-harmoniques a fait l'objet d'études à cause de son importance technique, soit qu'on recherche ce phénomène pour réaliser une division de fréquence<sup>1</sup>, soit qu'on l'évite à cause de son influence défavorable<sup>2, 3</sup>. Plus rares, et récents, sont les travaux<sup>4-8</sup> consacrés au cas des systèmes asservis. Ci-après nous présentons une étude de cette question pour les systèmes asservis par plus-ou-moins. Nous considérons (Figure 1) un système à retour unitaire (entrée  $e$ ; erreur  $\epsilon$ ) dont

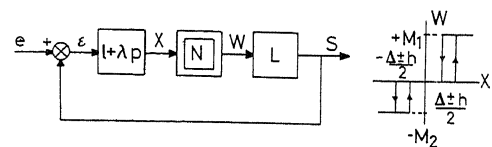


Figure 1

la branche directe comprend un correcteur  $1 + \lambda p$ , un organe non-linéaire  $N$  et une boîte  $L(p)$  linéaire de degré arbitraire. On sait<sup>9, 10</sup> que tous asservissements dont  $N$  est le seul organe non-linéaire, situé dans la chaîne directe, peuvent se ramener à ce type. La boîte  $N$  représente un élément par plus-ou-moins (ci-après appelé «relais») à caractéristique symétrique ( $w = M \text{ sign } x$ ) ou non ( $w = +M_1$  ou  $-M_2$ ), avec seuil  $\pm h/2$  et hystérésis  $\pm h/2$ .

On sait<sup>9,10</sup> que lorsqu'un tel système est forcé par une entrée périodique d'amplitude  $F$  et fréquence  $\omega_f$  il y a un seuil de synchronisation  $F_f$  qui dépend de  $\omega_f$ : une oscillation forcée fondamentale de fréquence  $\omega_f$  n'existe que si  $F > F_f$ . Le présent travail contribue à étudier le comportement du système dans la zone ( $F < F_f$ ), ombrée sur les Figures 3, 4, 5, où il ne peut y avoir d'oscillation fondamentale, en y étudiant la possibilité d'oscillations sous-harmoniques de fréquence  $\omega_f/n$ ,  $n$  étant entier. Il suppose connues les méthodes développées indépendamment par Hamel en France et Cypkin<sup>11</sup> en U.R.S.S. dont il existe des exposés élémentaires<sup>9,10</sup> indiquant les références originales.

### Sous-harmoniques dans le Cas Symétrique sans Seuil

( $\Delta = 0$ ,  $M_1 = M_2$ )

On sait<sup>11</sup> que le seuil de synchronisation  $F_f$  pour la fondamentale est donné lorsque  $\Delta = 0$  par

$$F_f = \left| \frac{h}{2} + \alpha \right|$$

$$= \frac{4M}{\pi} \left| V(\omega_f) + \frac{1}{3} V(3\omega_f) + \frac{1}{5} V(5\omega_f) + \dots + \alpha \right|$$

$V(\omega) = \text{Im } L(j\omega)$

Sakawa<sup>6</sup> et les présents auteurs<sup>7, 8</sup> ont montré facilement que le seuil de synchronisation  $F_n$  pour le sous-harmonique d'ordre  $n$  est de même donné par

$$F_n = \left| \frac{h}{2} + \alpha_n \right|$$

$$= \frac{4M}{\pi} \left| V\left(\frac{\omega_f}{n}\right) + \frac{1}{3} V\left(\frac{3\omega_f}{n}\right) + \frac{1}{5} V\left(\frac{5\omega_f}{n}\right) + \dots + \alpha_n \right| \quad (1)$$

Dans ces expressions  $\alpha_n$  (ou  $\alpha$ ) est l'ordonnée du point de fréquence  $n\omega_f$  (ou  $\omega_f$ ) du lieu de Cypkin<sup>9, 10</sup> du système.

D'où la technique suivante pour trouver des conditions nécessaires de production du sous-harmonique d'ordre  $n$ : (1) tracer la courbe  $F_f$  versus  $\log \omega_f$ , ombrer la zone sous-jacente; (2) la courbe  $F_n$  versus  $\log \omega_f$  s'obtient en la décalant vers la droite de  $\log_2 n$  octaves (1.6 octave pour  $n = 3$ , 2.3 pour 5, 2.8 pour 7, etc.); (3) alors le sous-harmonique d'ordre  $n$  peut s'observer dans la partie de la zone ombrée située au-dessus de la courbe  $F_n$ .

Nous croyons qu'il y a avantage à raisonner géométriquement dans le plan  $(\varepsilon, \dot{\varepsilon})$ , en utilisant le lieu de Hamel et non le lieu de Cypkin. On est amené à couper la droite de commutation  $D$  ( $\varepsilon + \lambda \dot{\varepsilon} = -h/2$ ) par le lieu  $\mathcal{E}_n$  déduit de la courbe fermée de représentation paramétrique  $e(t)$ ,  $\dot{e}(t)$  par la translation qui en amène le centre au point  $\omega_n = \omega_f/n$  du lieu de Hamel  $H$  (Figure 2). Ce point  $B_n$  a pour coordonnées

$$\alpha = \frac{4M}{\pi} \left[ V\left(\frac{\omega_f}{n}\right) + \frac{1}{3} V\left(\frac{3\omega_f}{n}\right) + \dots \right]$$

$$\beta = \frac{4M}{\pi} \left[ U\left(\frac{\omega_f}{n}\right) + U\left(\frac{3\omega_f}{n}\right) + \dots \right]$$

$U(\omega) = \text{Re } L(j\omega)$

Il n'y a sous-harmonique que si  $\mathcal{E}_n$  coupe  $D$  (et on montre facilement comme pour la fondamentale<sup>11</sup> qui seule est stable l'oscillation qui correspond à l'intersection la plus basse située).

Cette façon de raisonner présente les avantages suivants:

(1) L'inspection géométrique permet immédiatement une discussion qualitative, notamment de prévoir (i) les fonctions  $L(p)$  susceptibles de donner des sous-harmoniques — celles dont un segment important du lieu de Hamel, parcouru dans le sens des fréquences croissantes, s'éloigne de la droite de commutation — et (ii) le rôle favorisant de l'hystérésis.

(2) Sur la Figure 2, le point  $B_n$  décrit un lieu  $R_n$  facilement déduit de  $L(p)$ , et le point  $C_n[\varepsilon(t), \dot{\varepsilon}(t)]$  un lieu  $P_n$  qui se construit par composition de mouvements, comme expliqué ailleurs<sup>7, 8</sup> par les auteurs.  $C'_n$  représente une commutation de  $-$  à  $+$  et  $C_n$  est la commutation à mi-période. On voit que  $B_n C_n$  et  $B'_n C'_n$  sont parallèles et nécessairement de sens opposé, donc  $n$  est impair: il n'y a pas de sous-harmoniques d'ordre pair dans le cas symétrique.

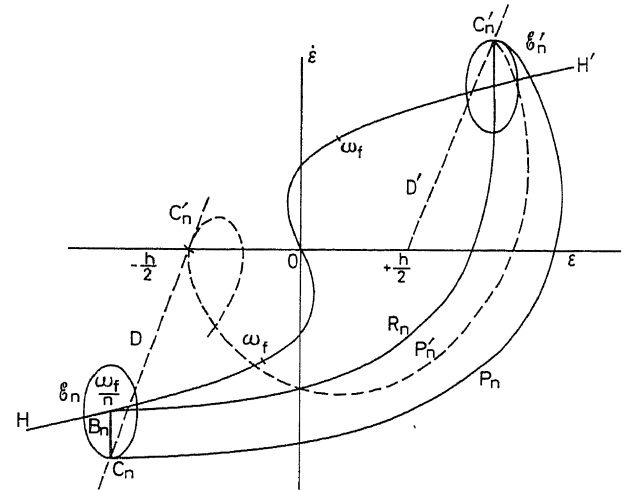


Figure 2. Production du sous-harmonique d'ordre  $n$ , cas symétrique. Le lieu  $P'_n$  montre une commutation prématurée

(3) Lorsqu'il y a un correcteur ( $\lambda > 0$ ) la méthode se généralise sans aucun nouveau calcul: le seuil  $F_n$  s'obtient en écrivant que  $\mathcal{E}_n$  est tangent à  $D$ . Dans le cas d'une entrée harmonique  $e(t) = F \sin \omega_f t$ , le lieu  $\mathcal{E}_n$  est l'ellipse

$$\omega_f^2 (\varepsilon - \alpha)^2 + (\dot{\varepsilon} - \beta)^2 = \omega_f^2 F^2$$

d'où l'expression explicite de  $F_n$

$$F_n^2 = \left| \left( \alpha + \frac{h}{2} \right)^2 + \frac{\beta^2}{\omega_f^2} - \frac{[(h/2 + \alpha)\lambda\omega_f^2 - \beta]^2}{\omega_f^2(1 + \lambda^2\omega_f^2)} \right| \quad (2)$$

dont (1) est un cas particulier.

(4) On possède une interprétation géométrique de la condition supplémentaire qu'il n'y ait pas de commutation avant la demi-période. Sur la Figure 2 la courbe  $P'_n$  coupe la droite de commutation de façon prématurée en  $C'_n$  avant la demi-période (pour  $t < n\pi/\omega_f$ ). On conçoit que les harmoniques d'ordre élevé, dont les courbes  $P'_n$  ont plusieurs points doubles, présentent facilement ce phénomène quand on agrandit  $\mathcal{E}_n$ , i.e. augmente l'amplitude de l'entrée. On en verra l'interprétation plus loin.

### Exemples

(a) Soit d'abord le système régulier  $L(p) = 1/p(1+p)$  ( $1+a$ ). La courbe  $F_f$  versus  $\log \omega_f$  a la forme indiquée par la courbe  $C_1$  des Figures 3 ou 4.

En l'absence d'hystérésis (Figure 3) des oscillations sous-

harmoniques de tous ordres sont possibles mais seulement dans des zones étroites du plan  $(\omega_f, F)$ . Le sous-harmonique d'ordre 3 apparaît dans la zone hachurée horizontalement; celui d'ordre 5 (et de même les suivants) n'apparaît que dans le tout petit triangle où la courbe  $C_5$  passe en-dessous de  $C_1$ , c'est-à-dire pour une entrée de très faible amplitude et dont la fréquence a justement été choisie extrêmement voisine de  $5\omega_0$ . Lorsque  $a = 0$  toutes les courbes  $C_n$  sont au-dessus de la courbe  $C_1$ : les sous-harmoniques n'apparaissent pas.

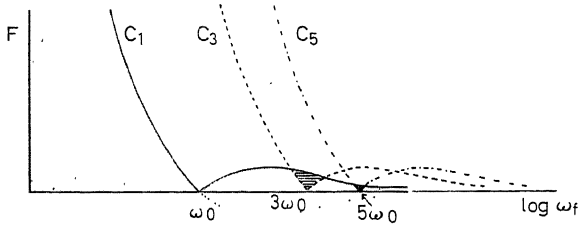


Figure 3. Sous-harmoniques d'un système régulier  $L(p) = 1/p(1+p)(1+ap)$  par plus-ou-moins sans hystérésis

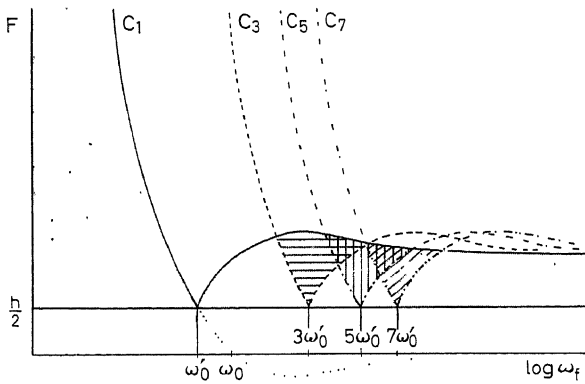


Figure 4. L'hystérésis facilite l'apparition des sous-harmoniques

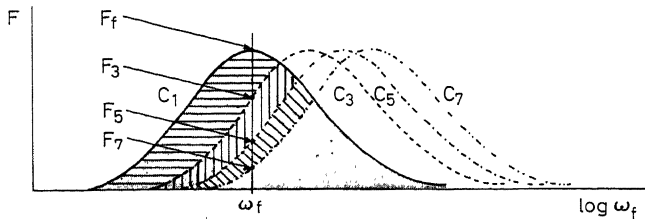


Figure 5. Sous-harmoniques du système particulier  $L(p) = p/(1+p)(1+ap)$

La présence d'hystérésis facilite l'apparition d'oscillations sous-harmoniques, car le lieu de Hamel décrit dans le sens des fréquences croissantes s'éloigne de la droite de commutation sur un segment plus long. On peut le vérifier aussi directement<sup>7, 8</sup>. La Figure 4 est relative au même cas ( $a = 0.1$ ) que la Figure 3. On constate de plus larges zones de possibilités pour les sous-harmoniques d'ordre 3 (hachures horizontales), 5 (verticales), 7 (obliques), etc., leur aire décroissant lorsque  $n$  augmente.

(b) Notre méthode permet d'étudier également le système particulier  $L(p) = p/(1+p)(1+ap)$  avec relais inversé. Les courbes  $F_f$  et  $F_n$  versus  $\log \omega_f$  sont alors celles de la Figure 5

( $a = 0.1$ ). Pour une fréquence  $\omega_f$  telle que celle indiquée sur la figure on peut avoir des sous-harmoniques de tous ordres, ceux d'ordres élevés étant obtenus pour de faibles valeurs de  $F$ . Pour des fréquences d'entrée plus hautes on ne peut produire que les sous-harmoniques d'ordre élevé.

Il est facile de retrouver directement ces résultats (Figure 6, où on a supposé  $a = 0$  pour rendre plus clair le mécanisme de génération des sous-harmoniques). On peut les expliquer par la forme du lieu de Hamel du système (la même que celle des

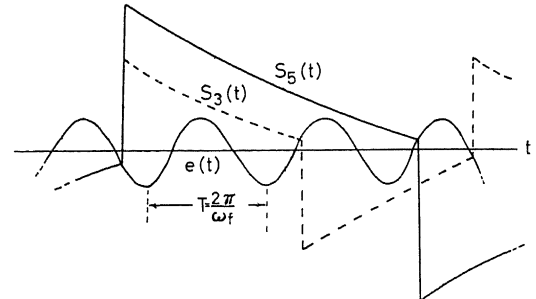
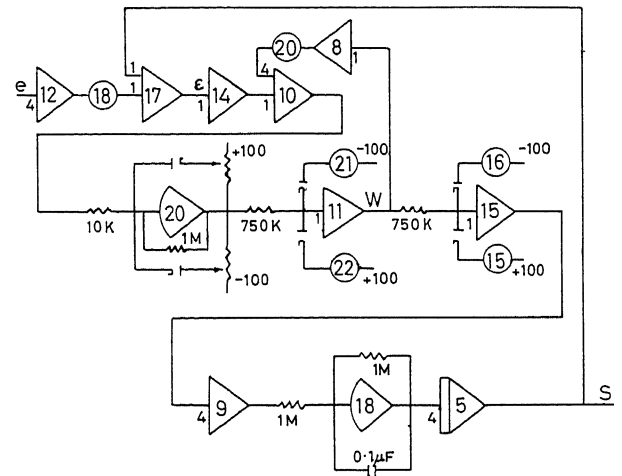


Figure 6. Oscillations sous-harmoniques d'ordres 3 et 5 du système particulier  $L(p) = p/(1+p)$



Figures 7. Simulation sur calculateur analogique du système  $L(p) = 4/p(1 + 0.1p)$

lieux de la Figure 10) car dans toute une bande  $0, \omega_m$  ce lieu s'éloigne de la droite de commutation, d'où la possibilité de tous sous-harmoniques.

#### Cas où Plusieurs Oscillations sont Simultanément Possibles

Il peut arriver qu'à une fréquence  $\omega_f$  deux sous-harmoniques d'ordres  $m$  et  $n$  ( $m < n$ ) soient possibles ( $F > F_m$  et  $F_n$ ). La théorie ci-dessus — comme la plupart des théories sur l'existence d'oscillations forcées — donnant seulement des conditions nécessaires d'existence, nous avons recouru à la simulation (montage du type de la Figure 7) pour savoir lequel s'observe effectivement, c'est-à-dire trouver les conditions également suffisantes.

(a) Si  $F_m < F_n$  ( $\omega < \omega_l$  sur la Figure 8), nous avons toujours observé le sous-harmonique d'ordre le moins élevé  $m$ .



(b) Si  $F_m > F_n$  (cas  $\omega > \omega_l$ ), la frontière présente une sorte d'hystérésis. (1) Lorsqu'on fait décroître l'amplitude d'entrée  $F$  on observe d'abord le sous-harmonique  $m$ , et le sous-harmonique  $n$  apparaît exactement pour  $F$  égal à la valeur  $F_m$  qu'indique la théorie, soit sur la courbe  $C_m$ . (2) Si au contraire, on franchit  $C_m$  de bas en haut on peut observer au-dessus de  $C_m$  l'un ou l'autre sous-harmonique, selon les conditions initiales du problème (cette dépendance est vraisemblablement complexe; expériment-

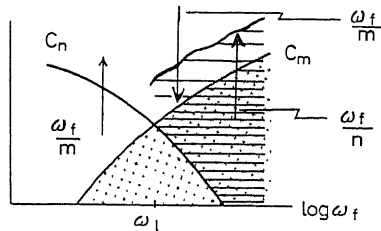


Figure 8. Frontière entre les oscillations d'ordres  $m$  et  $n$  quand on fait croître ou décroître l'amplitude d'entrée  $F$ : traînage de fréquence

talement on constate que la probabilité d'apparition du sous-harmonique d'ordre le plus élevé  $n$  décroît très vite quand on augmente  $F$ ; le passage au sous-harmonique d'ordre moins élevé  $m$  s'effectue par l'apparition d'une commutation prématurée qui fait disparaître le sous-harmonique d'ordre  $n$ . Ce «traînage de fréquence» s'observe sur les exemples des Figures 3 et 4. Il peut mettre en jeu les sous-harmoniques d'ordres 3 et 5, ou 5 et 7, etc.: dans ces cas il est spécialement facile à mettre en évidence sur le système de la Figure 4. Il existe également dans le cas où  $m = 1$ : le seuil de synchronisation du fondamental  $F_f$  présente lui aussi ce caractère d'hystérésis.

Au total, cette première approche du problème de la stabilité globale (conditions nécessaires et suffisantes) des oscillations forcées met en évidence les résultats suivants:

(1) Il existe un traînage de fréquence dans les systèmes par plus-ou-moins.

(2) A ce point de vue toutes les oscillations forcées (fondamentale et sous-harmoniques) jouent le même rôle. Donc dans la zone  $F > F_f$ , où on admettait habituellement qu'existe la fondamentale, les choses sont en réalité plus complexes et on peut parfois observer un sous-harmonique quand on opère en faisant croître  $F$ .

(3) Le phénomène de commutation prématurée était jusqu'ici considéré<sup>12</sup> comme une possibilité théorique guère rencontrée dans les systèmes réels: nous connaissons maintenant sa signification physique — la disparition d'un sous-harmonique au profit d'un autre d'ordre moins élevé — et comprenons ainsi pourquoi son intérêt réel n'a pu apparaître avant qu'on se penche sur le problème des sous-harmoniques.

### Cas d'un Relais à Caractéristique non Symétrique ou avec Seuil

Lorsque le relais a une caractéristique dissymétrique ( $w = +M_1$  ou  $-M_2$ ) les deux «demi-périodes»  $T_1$  et  $T_2$  du sous-harmonique ( $T_1 + T_2 = 2\pi n/\omega_f$ ) sont inégales. Pour les déterminer on peut<sup>5</sup> appliquer à la fréquence  $\omega_f/n$  la méthode proposée par Cypkin<sup>11</sup> pour la fondamentale. Cela conduit à centrer la courbe fermée  $\mathcal{C}_n$  aux différents points  $\omega_f/n$  des lieux de deux familles, et à en déduire par intersection avec la droite de commutation deux courbes du déphasage  $\phi$  versus  $\gamma = T_s/(T_1 + T_2)$ , qui se

coupent déterminant le  $\phi$  et le  $\gamma$  qui caractérisent le sous-harmonique d'ordre  $n$ .

Nous avons calculé et donnons en Figure 9 ces familles de lieux pour  $L(p) = 1/(1+p)(1+ap)$  avec  $a = 0,1$ ; les lieux  $\gamma = 0,5$  sont les lieux de Hamel pour l'oscillation symétrique traduits horizontalement de  $(M_1 - M_2)/2$ . La Figure 10 montre les familles de lieux pour le «système particulier» mentionné ci-dessus  $L(p) = p/(1+p)(1+ap)$  avec  $a = 0,1$ .

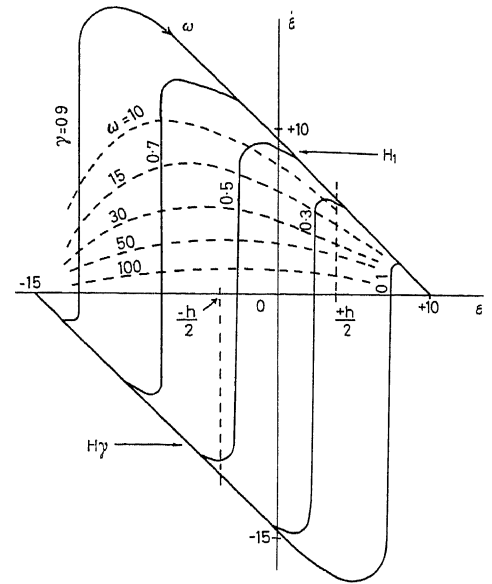


Figure 9. Familles de lieux de Hamel pour le système régulier  $L(p) = 1/(1+p)(1+0.1p)$  dans le cas d'un relais dissymétrique avec  $M_1 = 15$  et  $M_2 = 10$

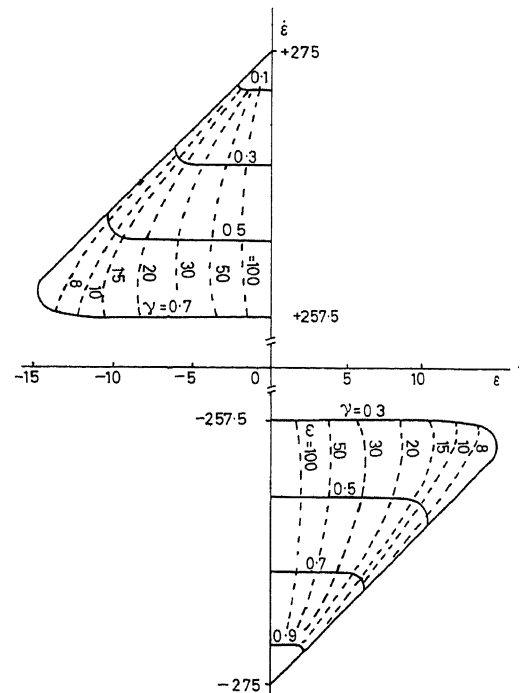


Figure 10. Familles de lieux de Hamel pour le système particulier  $L(p) = p/(1+p)(1+0.1p)$  dans le cas dissymétrique avec  $M_1 = 15$  et  $M_2 = 10$

Le cas d'un relais avec seuil peut se traiter de la même façon avec l'aide des familles de lieux de Cypkin définies au paragraphe 23 de Cypkin<sup>11</sup>.

Cette méthode résout théoriquement le problème. Dans le cas dissymétrique on note que cette fois les vecteurs  $B_n' C_n'$  ( $t = T_1 + T_2$ ) et  $B_n C_n$  ( $t = T_1$ ) pour  $\gamma$  donné ne sont plus nécessairement parallèles. On peut donc obtenir aussi des sous-harmoniques d'ordre pair. De plus l'inspection des lieux montre que la production de sous-harmoniques devient plus difficile à mesure que la dissymétrie s'accroît.

Malheureusement la discussion plus précise de l'influence des divers paramètres n'est guère possible à cause du grand nombre des courbes en jeu. Toutefois le problème se simplifie dans les cas où la fonction de transfert linéaire  $L(p)$  est telle qu'on connaisse *a priori* une relation entre  $T_1$  et  $T_2$ . Cela se produit notamment lorsque:

(1)  $L(p)$  possède une intégration, c.-à.-d. le pôle  $p=0$  (c'est le cas de la majorité des servomécanismes réels) — on a alors la relation  $T_1/T_2 = M_2/M_1$ ;

(2)  $L(p)$  possède le zéro  $p=0$  (exemple: le système particulier étudié en (3)) — on a alors  $T_1 = T_2$ .

Ci-après nous développons le calcul dans le premier cas, où nous montrerons que le problème de détermination d'une oscillation forcée, fondamentale ou sous-harmonique, peut se résoudre avec un lieu unique.

#### Calcul dans le cas où $L(p)$ Possède une Intégration

Soit d'abord le cas très simple où  $L(p) = 1/p (1 + Tp)$ . En intégrant

$$T\ddot{\varepsilon} + \dot{\varepsilon} = -KM_1 = -A$$

à partir des conditions initiales  $\varepsilon_0, \dot{\varepsilon}_0$  lors de la commutation de  $-$  à  $+$  il vient, posant  $r_i = \exp(-t_i/T)$  pour simplifier l'écriture:

$$\varepsilon = \varepsilon_0 - At + T(\dot{\varepsilon}_0 + A)(1 - r)$$

$$\dot{\varepsilon} = \dot{\varepsilon}_0 r - A(1 - r)$$

L'instant  $t_1$  pour lequel  $\varepsilon = -\varepsilon_0$  est donné par:

$$\varepsilon = -\varepsilon_0 = \varepsilon_0 - At_1 + T(\dot{\varepsilon}_0 + A)(1 - r_1)$$

On détermine ainsi:

$$\dot{\varepsilon}_0 = \frac{-2\varepsilon_0 + At_1 - AT(1 - r_1)}{T(1 - r_1)} \quad (3)$$

et

$$\dot{\varepsilon}(t_1) = \dot{\varepsilon}_0 r_1 - A(1 - r_1)$$

$$\dot{\varepsilon}(t_1) = \dot{\varepsilon}_0 + (2\varepsilon_0 - At_1)/T$$

$$\dot{\varepsilon}(t_1) = \frac{r_1(-2\varepsilon_0 + At_1) - AT(1 - r_1)}{T(1 - r_1)} \quad (4)$$

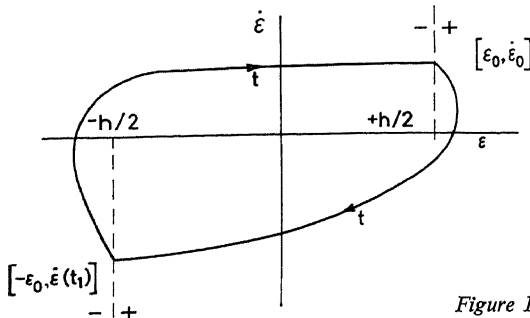


Figure 11

De la même façon, l'intégration de

$$T\ddot{\varepsilon} + \dot{\varepsilon} = +KM_2 = B$$

à partir des conditions  $\varepsilon_1 = -\varepsilon_0$  et  $\dot{\varepsilon}_1 = \dot{\varepsilon}(t_1)$ , donne

$$\varepsilon = \varepsilon_1 + Bt + T(\dot{\varepsilon}_1 - B)(1 - r)$$

$$\dot{\varepsilon} = \dot{\varepsilon}_1 r + B(1 - r)$$

L'instant  $t_2$  auquel  $\varepsilon$  est revenu en  $\varepsilon_0$  est donné par

$$2\varepsilon_0 = Bt_2 + T(\dot{\varepsilon}_1 - B)(1 - r_2) \quad (5)$$

On détermine ainsi

$$\dot{\varepsilon}_1 = \dot{\varepsilon}(t_1) = \frac{2\varepsilon_0 - Bt_2 + BT(1 - r_2)}{T(1 - r_2)}$$

En prescrivant alors, pour la périodicité,  $\dot{\varepsilon}(t_2) = \dot{\varepsilon}_0$  (ce qui revient à appliquer à la droite de commutation  $D'$  de  $-$  à  $+$  la méthode générale de la transformation ponctuelle<sup>14, 15</sup>), on trouve que les deux demi-périodes  $T_1 = t_1$  et  $T_2 = t_2$  doivent satisfaire la relation annoncée

$$T_1 M_1 = T_2 M_2$$

qui détermine directement le rapport  $\delta = T_2/T_1$  des demi-périodes:

$$\delta = M_1/M_2 \quad (6)$$

Les conditions de périodicité sont donc les relations (3) à (5) qui, avec (6), déterminent  $\varepsilon_0$  et  $\dot{\varepsilon}_0$ . La résolution de ce système transcendant peut se faire soit graphiquement, soit pour plus de précision avec calculateur (machine ANALAC, ou par itération avec un calculateur digital). Pour chaque valeur de  $\delta$  (qui cette fois est connu) on construit ainsi (Figure 12, où  $T = 0,1$ ) un lieu unique  $\dot{\varepsilon}_0$  versus  $\varepsilon_0$  gradué en  $(T_1 + T_2)$ , qui donne directement l'oscillation libre et le sous-harmonique par intersection, avec la droite de commutation  $D'$  de  $-$  à  $+$ , respectivement de ce lieu et de la courbe fermée  $(e, \dot{e})$  (ellipse si l'entrée est harmonique) centrée en son point  $\omega = 2\pi/n(T_1 + T_2)$ .

La méthode est exactement la même pour des fonctions  $L(p)$  avec intégrations plus compliquées, telles que  $1/p(1 + Tp)(1 + T'p)$ . Les seconds membres des équations (3) à (5), obtenus

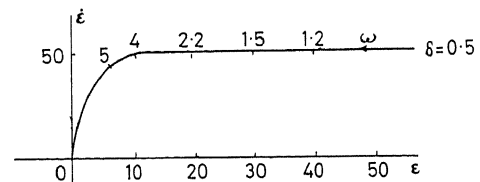


Figure 12. Lieu de Hamel unique pour le système régulier simple  $L(p) = 4/p(1 + 0.1p)$  dans le cas dissymétrique  $M_1 = 6.25, M_2 = 12.5$  ( $\delta = 0.5$ )

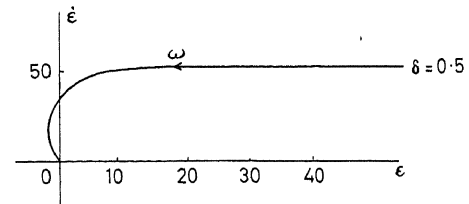


Figure 13. Lieu de Hamel unique pour le système régulier  $L(p) = 4/p(1 + p)(1 + 0.1p)$  dans le cas dissymétrique  $M_1 = 6.25, M_2 = 12.5$  ( $\delta = 0.5$ )

avec l'aide d'expressions connues<sup>16</sup>, ont une écriture plus compliquée, la quantité  $\delta$  est toujours donnée directement par l'équation (6) et pour chaque  $\delta$  on a un lieu unique gradué en  $\omega$  (Figure 13) à partir duquel les sous-harmoniques se déterminent comme dans le cas symétrique. Notamment, les seuils de synchronisation s'obtiennent en traçant *versus*  $\log \omega_f$  l'abscisse du lieu par rapport à la droite de commutation  $D'$  et en décalant cette courbe parallèlement à l'axe des fréquences.

La présente étude a été rendue possible grâce à un généreux don du Conseil des Recherches pour la Défense (Canada).

## Références

- <sup>1</sup> GROSZKOWSKI, J. Frequency division. *Proc. Inst. Radio Engrs* 18, (11), (1930)
- <sup>2</sup> FALLOU, J. Démultiplieur de fréquence statique. *Rev. Gén. Élect.* 19 (1926), 987
- <sup>3</sup> BRENNER, E. Subharmonic response of the ferroresonant circuit with coil hysteresis. *Commun. & Electron. Trans. Amer. Inst. elect. Engrs* (1956), 450
- <sup>4</sup> WEST, J., et DOUCE, J. The mechanism of subharmonic generation in a feedback system. *Proc. Instr. elect. Engrs* 103 (B) (1954) 569
- <sup>5</sup> YAMAGUCHI, J., NISHIMURA, M., FUJII, K., et MARUHASHI, T. On the subharmonic oscillation of the relay servomechanism. *Automatic and Remote Control*, Vol. 1, p. 398. 1961 Londres; Butterworths
- <sup>6</sup> SAKAWA, Y. Subharmonic oscillations in relay-control systems. *Automatic and Remote Control*, Vol. 1, p. 404. 1961. Londres; Butterworths
- <sup>7</sup> GILLE, J. C. et PAQUET, J. G. Subharmonic oscillations in on-off control systems. *Amer. Inst. elect. Engrs, Conference Paper* CP-61-937, 1961
- <sup>8</sup> GILLE, J. C. et PAQUET, J. G. Oscillations sous-harmoniques dans un asservissement par plus-ou-moins. *Automatisme* 7 (1) (1961) 5-10; *Bull. Acad. Sci. de Pologne (Sc. techn.)* (5) (1962) 279
- <sup>9</sup> GILLE, J. C., PELEGRIN, M., et DECAULNE, P. *Feedback Control Systems*. pp. 390, 446. 1959. New York; McGraw-Hill
- <sup>10</sup> GILLE, J. C., DECAULNE, P., et PELEGRIN, M. *Méthodes modernes d'étude des systèmes asservis*, pp. 164, 252. 1960. Paris; Dunod
- <sup>11</sup> CYPKIN, JA. Z. *Teorija relejnykh sistem avtomatičeskogo regulirovaniya*. pp. 154, 206, 238. 1955. Moscou; Gostehizdat
- <sup>12</sup> NEJMARK, JU. I. O periodičeskikh režimakh i ustojčivosti relejnykh sistem. *Automat. Telemekh.* 14 (5) (1953) 556
- <sup>13</sup> GILLE, J. C. A propos des oscillations libres des asservissements par plus-ou-moins. *Automat. strumetaz.* 8 (1) (1960) 3-10
- <sup>14</sup> ANDRONOV, A., VITT, A., et KHAIKIN, S. *Teorija kolebanij*, p. 328, 1959. Moscou; Fizmatgiz
- <sup>15</sup> SOLODOVNIKOV, V. *Grundlagen der selbsttätigen Regelung* (traduit du russe, Oldenbourg, Munich et Technik, Berlin 1959) p. 817
- <sup>16</sup> HOPKIN, A., et WANG, P. Further studies of relay-type feedback control systems designed for random inputs. *Automatic and Remote Control*, Vol. 1, p. 369. 1961. Londres; Butterworths

# Synthesis of Control Systems Operating Linearly for Small Signals and Approximately 'Bang-Bang' for Large Signals

E. V. PERSSON

## Summary

A non-linear controller operating linearly in the small-signal range and approximately 'bang-bang' in the large-signal range is proposed. Its application to a process comprising  $n$  integrators with limited  $n$ th derivative of the controlled output is treated. The non-linear feedback functions of the controller are given such a form that the response in principle is independent of the signal amplitude. The possibilities of giving the system a bang-bang character for large control deviations are investigated.

Analogue-computer studies have been made on systems with limited second, third and fourth derivatives. These studies show that very satisfactory results are achieved for the large-signal response, which deviates only insignificantly from the bang-bang response.

The proposed type of controller is very simple in application, since it is only necessary to make non-linear the normal feedbacks in a conventional controller.

## Sommaire

L'auteur propose un régulateur non-linéaire qui fonctionne linéairement dans la gamme des petits signaux et approximativement «bang-bang» dans la gamme des grands signaux. Il traite de son application à un processus qui comprend  $n$  intégrateurs avec limitation de la  $n$ ième dérivée de la grandeur réglée.

On donne aux asservissements non linéaires du régulateur une forme telle qu'en principe, la réponse soit indépendante de l'amplitude du signal.

On examine les possibilités de donner au système un caractère «bang-bang» pour de grands écarts de réglage.

Des études au calculateur analogique ont été faites sur des systèmes avec limitation de la seconde, troisième et quatrième dérivées. Ces études montrent qu'on obtient une réponse très satisfaisante pour de grands signaux, réponse qui ne diffère que d'une façon insignifiante de la réponse bang-bang.

Le type de régulateur proposé est très simple à réaliser en pratique de fait qu'il est seulement nécessaire de rendre non-linéaires les réactions normales d'un régulateur classique.

## Zusammenfassung

Es wird ein nichtlinearer Regler vorgeschlagen, der bei kleinen Signalamplituden ein lineares Verhalten, bei großen Signalamplituden angenähert Zweipunktverhalten aufweist. Seine Anwendung für eine Regelstrecke, die aus  $n$  Integratoren mit beschränkter  $n$ ter Ableitung der Regelgröße besteht, wird behandelt. Die nichtlinearen Rückführungen sind so ausgeführt, daß die Übergangsfunktion im Prinzip von der Signalamplitude unabhängig ist. Die Möglichkeiten, dem Regelsystem bei großen Regelabweichungen Zweipunktverhalten zu geben, werden betrachtet.

Untersuchungen von Systemen mit beschränkter zweiter, dritter und vierter Ableitung am Analogrechner zeigen, daß sich sehr befriedigende Ergebnisse erzielen lassen. Die Übergangsfunktion bei großen Signalamplituden weicht nur geringfügig vom Zweipunktverhalten ab.

Der vorgeschlagene Regler läßt sich in der Praxis leicht verwirklichen, da es nur notwendig ist, die normalen Rückführungen eines konventionellen Reglers nichtlinear zu machen.

## Introduction

Conventional control systems designed as linear systems operate satisfactorily within the limits of linearity, and the desired response and accuracy can be achieved by means of correct design for which effective methods of synthesis exist.

If, however, a system of this kind is subjected to such large changes in the command input or such large disturbances that limiting is reached, the response may be unsatisfactory and many systems will even become unstable. Thus a suitable compromise must be made between large-signal and small-signal response, a compromise which may mean that neither will be fully satisfactory.

If, on the other hand, one designs the system for the fastest possible response with the control effort available, this will lead to a bang-bang system. Such a system will be very complicated, however, and difficult to apply if the process is described by a differential equation of higher order than the second. Furthermore the bang-bang response is not desirable for small control errors, since very high-frequency oscillations will appear at steady state.

It has been proposed<sup>1</sup> that the non-linear controller should be disconnected when the error has been corrected and that the steady state should be maintained by connecting up a new controlling device. It is hardly likely, however, that such a system would operate satisfactorily for small disturbances occurring frequently.

This paper presents a system which operates mainly bang-bang for large control errors but linearly for small errors. The transition between the two modes of operation takes place smoothly without any switching devices. The author has aimed at a system which is simple to apply, and for this reason the requirements as to an exact bang-bang response have been waived.

This paper is limited to the case where the control loop comprises a number ( $n$ ) of integrators and where the input to the first integrator is limited to its absolute value, i.e. the  $n$ th derivative of the controlled output is limited.

## The Problem

Let us start with a non-stabilized control system according to Figure 1, where it is assumed that the process is represented by  $n$  integrators.

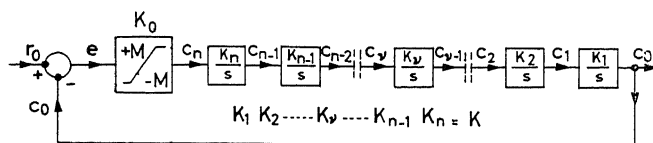


Figure 1. Non-stabilized system under study

$r_0$  = command input;  $c_0$  = controlled output;  $M$  = limiting level for  $c_n$

The following notation is used:

- $r_0$  = Command input
- $r_v = d^v r_0 / d t^v = v$ th derivative of the command input
- $c_0$  = Controlled output
- $c_v$  = Variable equal to constant  $\cdot d^v c_0 / d t^v$  (see Figure 1)
- $c_n$  = Input to first integrator
- $M$  = Limiting value for  $c_n$  according to:  $-M < c_n < M$
- $e$  = Input to the non-linear function representing the limitation
- $K_v (v = 1 \dots n)$  = Gain of the integrators
- $K = \prod_{v=1}^{v=n} K_v$
- $K_0$  = Gain of the non-linear function within the linear range
- $t$  = Time
- $s$  = Laplace operator

It is now desired to design a practicable non-linear controller, operating linearly for small signals and functioning satisfactorily even for large signals. The bang-bang response is taken as the ideal for large-signal response.

The command input is assumed to have the form

$$r_0 = b_0 + b_1 t + b_2 t^2 + \dots b_{n-1} t^{n-1} \quad (1)$$

where  $b_0 \dots b_{n-1}$  are constants. As is apparent, it has been assumed that:

$$r_n = \frac{d^n r_0}{d t^n} = 0 \quad (2)$$

### Normalizing the Equations to Reduce the Number of Parameters

The following variables are now introduced instead of the original ones:

$$\tau = \sqrt[n]{K} \cdot t \quad (3)$$

$$p = \frac{d}{d\tau} = \frac{1}{\sqrt[n]{K}} \frac{d}{dt} = \frac{1}{\sqrt[n]{K}} s \quad (4)$$

$$x_0 = \frac{c_0 - r_0}{M} \quad (5)$$

$$x_1 = \frac{dx_0}{d\tau} = \frac{c_1 K_1 - r_1}{M \sqrt[n]{K}} \quad (6)$$

$$\dots$$

$$x_v = \frac{d(x_{v-1})}{d\tau} = \frac{c_v K_1 K_2 \dots K_{v-1} K_v - r_v}{M [\sqrt[n]{K}]^v} \quad (7)$$

$$\dots$$

$$x_n = \frac{d(x_{n-1})}{d\tau} = \frac{c_n K_1 K_2 \dots K_{n-1} K_n}{M [\sqrt[n]{K}]^n} = \frac{c_n}{M} \quad (8)$$

$$y = \frac{K_0}{M} e \quad (9)$$

With the new variables the block diagram is shown in Figure 2.  $x_0$  is the control error (disregarding scale factor) [see eqn (5)];  $x_v (v = 1 \dots n)$  is the  $v$ th derivative of  $x_0$  with regard to  $\tau$ ; and the limiting value for  $x_n$  is unity.

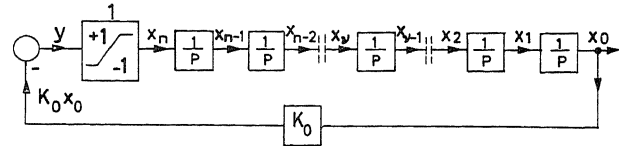


Figure 2. Normalized system

$x_0$  = control error in relative scale =  $c_0 - r_0/M$ ;  $x_v = d^v x_0 / d \tau^v$ ; where  $\tau = \sqrt[n]{K} t$

Since  $d^n r_0 / d t^n = 0$  according to eqn (2), all the variables  $x_v$  at steady state will assume the value 0, and the behaviour of the system, according to Figure 1, can be studied instead in the system shown in Figure 2 by allowing the variable  $x_v$  to assume the initial value

$$-\frac{(r_v)_{t=0}}{M [\sqrt[n]{K}]^v}$$

### Introduction of a Non-linear Controller

A non-linear controller is now introduced into the system shown in Figure 2. The variable  $y$  is determined by the control error  $x_0$  and its derivatives  $x_1, x_2 \dots x_{n-1}$  with regard to the normalized time  $\tau$ . In order to make this controller practicable,  $y$  is assumed to be the sum of the single-variable non-linear functions  $f_0(x_0), f_1(x_1) \dots f_{n-1}(x_{n-1})$  as shown in Figure 3.

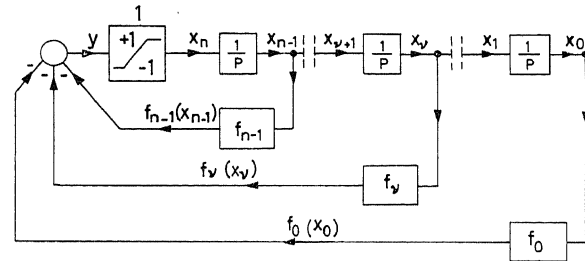


Figure 3. Introduction of non-linear controller with feedback functions  $f_v$

### The Large-signal Response of the System. Form of the Non-linear Feedback Functions

With very large signals the non-linear function representing the limitation in Figure 3 is greatly overloaded with the output  $x_n = \pm 1$  dependent on whether  $y$  (the switching function) is positive or negative.

If it is now assumed that it is possible to determine the non-linear feedback functions  $f_v(x_v)$  so that the response will be that desired for a certain large signal level, e.g. for a certain initial value of  $x_0$ , it is natural to require the same response for other large signal levels, though with altered amplitude and time scales. Which types of non-linear functions fulfil this requirement are investigated here. A further time transformation is now made for this purpose as follows:

$$\tau' = \beta \tau \quad (10)$$

$$\frac{d}{d\tau'} = q$$

where  $\beta$  is a positive constant and  $q$  is a new operator determined from

$$p = \beta q \quad (11)$$

The block diagram in Figure 3, due to the transformation (11), becomes that shown in Figure 4, which in turn can be redrawn by means of amplitude transformations giving that shown in Figure 5.

Comparing Figure 5 with Figure 3, it is apparent that the systems are exactly the same except that  $x_v$  is replaced by  $\beta^{n-v} x_v$  and that the switching function  $y$  is replaced by  $y'$ . If the non-linear feedback functions are now selected so that  $y'$  will also be a positive constant times  $y$  (the constant may be a function of  $\beta$ ), then  $\beta^{n-v} x_v$  as a function of  $\tau'$  and  $x_v$  as a function of  $\tau$  will be identical for the systems shown in Figures 5 and 3.

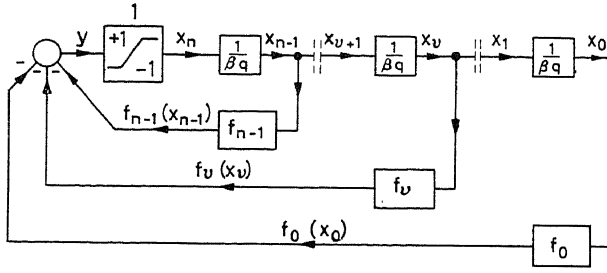


Figure 4. System with modified time-scale

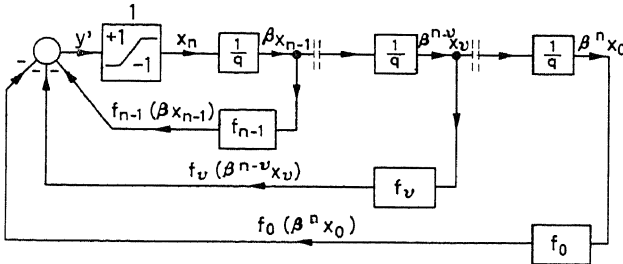


Figure 5. System with modified time- and amplitude-scale

The condition for this is consequently

$$\begin{aligned} y' &= \phi(\beta) y \\ f_v(\beta^{n-v} x_v) &= \phi(\beta) f_v(x_v) \end{aligned} \quad (12)$$

where  $\phi(\beta)$  is an arbitrary, positive function of  $\beta$ .

The solution of (12) is

$$|f_v(x_v)| = k_v |x_v|^{\frac{B}{n-v}}$$

Since  $f_v(x_v)$  must be an odd function of  $x_v$ , the final solution will be

$$f_v(x_v) = k_v \frac{x_v}{|x_v|} |x_v|^{\frac{B}{n-v}} \quad (13)$$

In eqn (13)  $k_v$  is a constant that can be freely selected for each  $v$  value, whereas  $B$  is an arbitrary constant that must have the same value for all values of  $v$ .

Now that the form of the function is given according to eqn (13), the practical synthesis work can be accomplished without difficulty by aid of an analogue computer. This work may be limited to one large amplitude, and suitable values for the  $n$  constants  $k_v$  can be determined. When a desired response has been obtained, the same response for other large amplitudes will automatically be obtained, although in another time scale.

An attempt is now made, however, to determine the constants  $k_v$  so that the response will approach the bang-bang response as closely as possible.

#### Determining the Constants $k_v$ to Approach Bang-bang Response

The bang-bang response comprises a series of time intervals with the duration  $\tau_1, \tau_2, \dots, \tau_n$  according to Figure 6, with  $x_n$  varying between  $+1$  and  $-1$ . The numbering has been chosen so that  $\tau_1$  refers to the last interval. At the end of this interval ( $\tau = \tau_s$ ) the state of equilibrium, characterized by the fact that all the  $x_v$  values are zero, is attained.

In order to simplify the calculations a new normalized, but reversed time  $\vartheta$ , is introduced according to

$$\vartheta = \tau_s - \tau \quad (14)$$

which means that the final state for  $\tau = \tau_s$  becomes the initial state for  $\vartheta = 0$ , namely:

$$x_0 = x_1 = \dots = x_{n-1} = 0 \text{ for } \vartheta = 0 \quad (15)$$

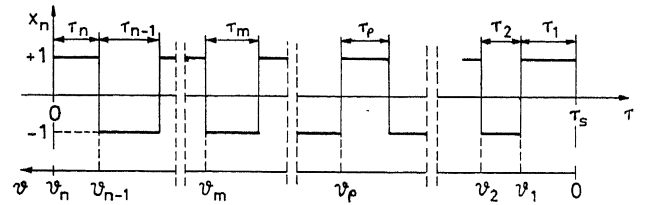


Figure 6. Bang-bang response

The variables  $x_v$  are now related by the following equations:

$$\begin{aligned} x_v &= - \int_0^{\vartheta} x_{v+1} d\vartheta \\ v &= 0 \dots n-1 \end{aligned} \quad (16)$$

Starting with  $x_n$  according to Figure 6, interpreted as a sum of unit step functions, and applying successively eqn (16) gives the following value for  $x_v$  at  $\vartheta = \vartheta_m$ :

$$x_{vm} = (-1)^{n-v} \frac{1}{(n-v)!} \left[ \vartheta_m^{n-v} + 2 \sum_{\rho=1}^{n-v} (-1)^\rho (\vartheta_m - \vartheta_\rho)^{n-v} \right] \quad (17)$$

At this instant  $x_n$  should change sign, which means that the switching function  $y$  should be equal to zero.

The switching function has the form:

$$y = - \sum_{v=0}^{n-1} f_v(x_v) = - \sum_{v=0}^{n-1} k_v \frac{x_v}{|x_v|} |x_v|^{\frac{B}{n-v}} \quad (18)$$

Inserting for  $x_v$  in eqn (18) values  $x_{vm}$  according to eqn (17) and setting  $y$  equal to zero, yields the following condition for bang-bang response:

$$\sum_{v=0}^{n-1} k_v (-1)^{n-v} \left[ \frac{1}{(n-v)!} \right]^{\frac{B}{n-v}} \cdot \frac{\Phi_m(\vartheta)}{|\Phi_m(\vartheta)|} \cdot |\Phi_m(\vartheta)|^{\frac{B}{n-v}} = 0 \quad (19)$$

$$m = 1, 2, \dots, n-1$$

where

$$\Phi_m(\vartheta) = \vartheta_m^{n-v} + 2 \sum_{\rho=1}^{n-v} (-1)^\rho (\vartheta_m - \vartheta_\rho)^{n-v}$$

Equations (19) must be satisfied, by selecting the constants  $k_v$  and  $B$ , for all values of  $\vartheta_1, \vartheta_2 \dots \vartheta_{n-1}$  if the system is to operate bang-bang irrespective of the initial condition. This is not possible, however, except in the simple case  $n = 2$ . One is therefore restricted to requiring bang-bang response only for a step input.

In this case the following applies, according to Burmeister<sup>2</sup>, for the relative duration of the intervals in Figure 6:

$$\frac{\vartheta_r}{\vartheta_n} = \sin^2 \frac{r\pi}{2n}$$

for  $r = 1, 2 \dots n-1$ .

From this:

$$\vartheta_m = \vartheta_n \sin^2 \frac{m\pi}{2n} \quad (20)$$

$$\vartheta_\rho = \vartheta_n \sin^2 \frac{\rho\pi}{2n} \quad (21)$$

Using the relationships (20) and (21) and introducing the following function  $A$  of  $n, v$  and  $m$ :

$$A(n, v, m) = 1 + 2 \sum_{\rho=1}^{m-1} (-1)^\rho \left[ 1 - \frac{\sin^2 \frac{\rho\pi}{2n}}{\sin^2 \frac{m\pi}{2n}} \right]^{n-v} \quad (22)$$

then eqn (19) can be transformed into

$$\sum_{v=0}^{n-1} k_v (-1)^{n-v} \left[ \frac{1}{(n-v)!} \right]^{\frac{B}{n-v}} \frac{A(n, v, m)}{|A(n, v, m)|} |A(n, v, m)|^{\frac{B}{n-v}} = 0 \quad (23)$$

$m = 1, 2, \dots, n-1$

The system of eqns (23) contains  $n-1$  equations, which must be satisfied by selection of the  $n$  constants  $k_0, k_1 \dots k_{n-1}$  and the constant  $B$ , and this is always possible.  $B$  and one of the constants  $k_v$  can be arbitrarily selected. The system of eqns (23) determines, however, for a given value of  $B$  the relationship between the  $n$  constants  $k_v$ . The calculation has been performed for  $n = 2, 3$  and  $4$ , the result being as follows:

$$\frac{k_1}{k_0} = \left( \frac{1}{2} \right)^{B/2} \quad (24)$$

$$\frac{k_1}{k_0} = \frac{2^{B/2}}{6^{B/3}} \frac{11^{B/3} + 1}{2} \quad (25)$$

$$\frac{k_2}{k_0} = \frac{1}{6^{B/3}} \frac{11^{B/3} - 1}{2}$$

When synthesizing according to eqns (24) to (26), the switching function will be zero at all those instants necessary for bang-bang step response. This does not necessarily mean, however, that the system must operate bang-bang, since it is possible that the switching function will pass through zero at other instants as well. An investigation of the derivatives of the switching function at the zero points suggests that the step response should be bang-bang for  $n = 2$  and  $n = 3$  irrespective of the value of  $B$ . For  $n = 4$ , the step response should be

bang-bang for  $B \leq 2$ , but not for  $B > 2$ . When  $B > 2$ , the first switching is correct, as is also the beginning of the second one. The second switching is abnormal, however, with a number of rapid oscillations about zero taking place over a short interval.

If the results are to be of practical use, it is naturally also necessary for the system to be stable; this is not certain even if the system has a bang-bang step response.

$$\frac{k_1}{k_0} = \frac{6^{B/2} \{ [4(\sqrt{2}+1)^4 - 1]^{B/4} + 1 \} (\sqrt{2}-1)^B + 2 \left( \frac{1}{2} \right)^{B/4}}{24^{B/4} \left[ \left( \frac{\sqrt{2}-1}{\sqrt{2}} \right)^{B/3} + (\sqrt{2}-1)^B \right]} \quad (26)$$

$$\frac{k_2}{k_0} = \frac{2^{B/2} [4(\sqrt{2}+1)^4 - 1]^{B/4} - 1}{24^{B/4} \cdot 2}$$

$$\frac{k_3}{k_0} = \frac{1}{24^{B/4}} \frac{\{ [4(\sqrt{2}+1)^4 - 1]^{B/4} + 1 \} \left( \frac{\sqrt{2}-1}{\sqrt{2}} \right)^{B/3} - 2 \left( \frac{1}{2} \right)^{B/4}}{2 \left[ \left( \frac{\sqrt{2}-1}{\sqrt{2}} \right)^{B/3} + (\sqrt{2}-1)^B \right]}$$

#### Modifying the Non-linear Functions to give Linear Operation for Small Signals

If the system in Figure 3 is to operate linearly within the small-signal range, the following must apply here:

$$f_v(x_v) = a_v x_v \quad (27)$$

where  $a_v$  are constants selected by means of synthesis methods for linear systems.

The following applies for large signals [see eqn (13)]:

$$f_v(x_v) = k_v \frac{x_v}{|x_v|} |x_v|^{\frac{B}{n-v}} \quad (28)$$

The constants  $k_v$  are selected according to eqns (24)–(26) or from computer studies. Depending on whether  $B/n - v$  is less than, equal to, or larger than unity, the functions will have the appearance shown in Figure 7 (a), (b) or (c). For the case  $B/n - v = 1$ ,  $k_v$  has been selected equal to  $a_v$ . In this case the function is the same in both ranges. For the other cases, where  $B/n - v \neq 1$ , the full-drawn curves in Figure 7 are selected.

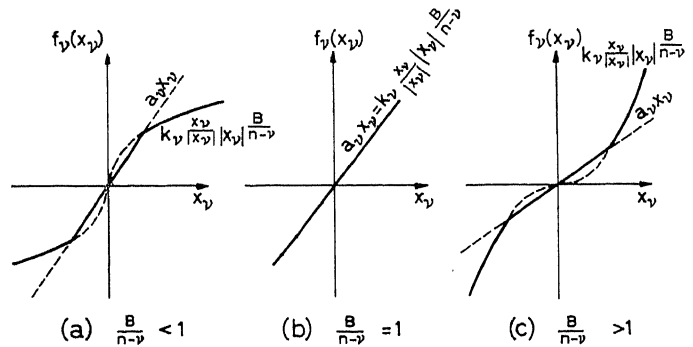


Figure 7. Proposed non-linear feedback function

When making this choice of the functions, the system will operate linearly for small signals and approximately bang-bang for such large signals that the linear range will be negligibly small. The transition between these two modes of operation takes place gradually without any need for switching devices.

When selecting the non-linear functions according to eqn (28), the constant  $B$  in the exponent can be given any positive value. In the first place, however, a whole number ought to be selected within the range  $1 \leq B \leq n$ .

With this choice, one of the functions will be linear, namely  $f_{n-B}(x_{n-B})$  which is equal to  $k_{n-B}x_{n-B}$ . The constant  $k_{n-B}$  is determined by

$$k_{n-B} = a_{n-B}$$

For practical reasons  $B = 1$  is preferred, provided that this is consistent with satisfactory response.

The choice of appropriate values for  $B$  and for the constants  $k_v$  is best made with the aid of an analogue computer. Such an investigation has been started but not yet completed, and it is thus only possible to report the results obtained from preliminary investigations, which were intended to establish what possibilities exist to improve a system by introducing a non-linear controller of the type proposed.

### Preliminary Analogue Computer Studies

Preliminary investigations have been performed on systems with two, three and four integrators ( $n = 2, 3$  and  $4$ ). Some of the results are presented here.

The synthesis in the linear range (selection of the constants  $a_v$ ) has been made according to a method described in the work by Kessler<sup>3</sup>. This synthesis method will yield a system with a step response overshoot of about 5 per cent.

#### Case $n = 2$

In this case the system operates bang-bang for all initial conditions disregarding the small deviation due to the linear range in the non-linear functions.

Some responses are presented in Figure 8. The feedback functions are as follows. In the linear range:

$$a_0 = 2; \quad f_0(x_0) = 2x_0$$

$$a_1 = 2; \quad f_1(x_1) = 2x_1$$

In the non-linear range:

$$B = 1$$

$$k_1 = a_1 = 2 \quad f_1(x_1) = 2x_1$$

$$k_0 = 2\sqrt{2} \quad [\text{from eqn (24)}] \quad f_0(x_0) = 2\sqrt{2} \frac{x_0}{|x_0|} |x_0|^{1/2}$$

The step response is shown in Figure 8 (b) and (d) for two different amplitudes of the step. For comparison the corresponding step responses with a linear controller are shown in Figure 8 (a) and (c). As is apparent, the non-linear controller operates almost ideally bang-bang for large steps.

#### Case $n = 3$

In this case the system is not stable with  $B = 1$  and the constants  $k_v$  selected according to eqn (25). With  $B = 3$ , however, a system operating practically bang-bang is obtained. A few responses are shown in Figures 9 and 10 with the following feedback functions [not true for Figure 9 (d)].

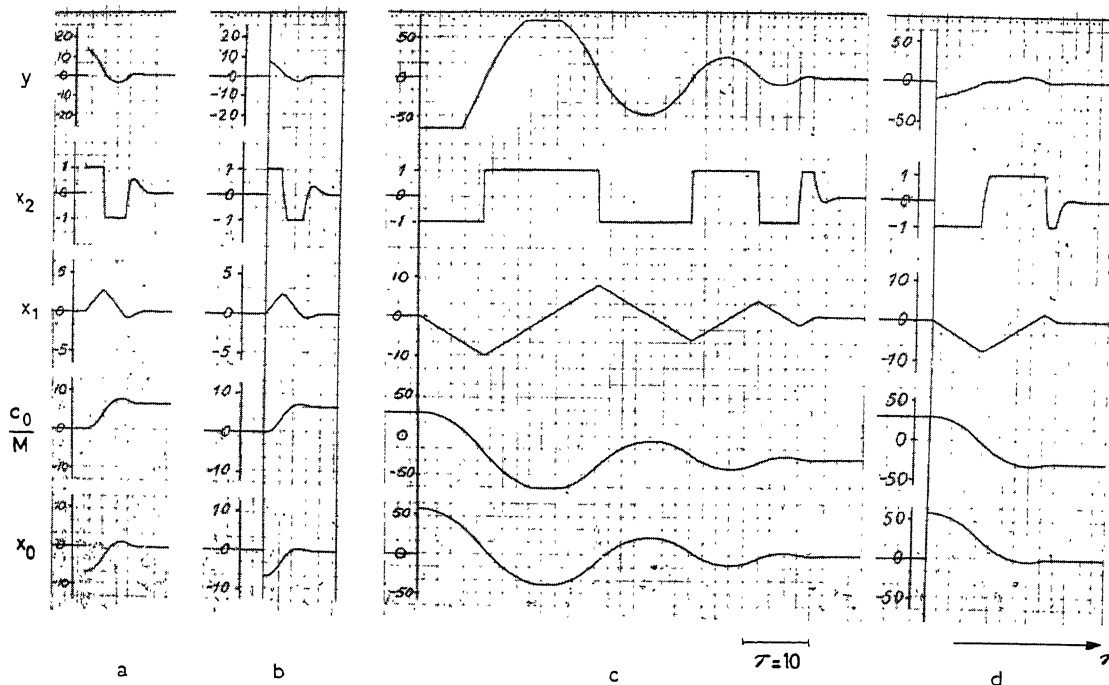


Figure 8. Step response for system with  $n = 2$ : (a) Linear controller; (b) non-linear controller; (c) linear controller; (d) non-linear controller



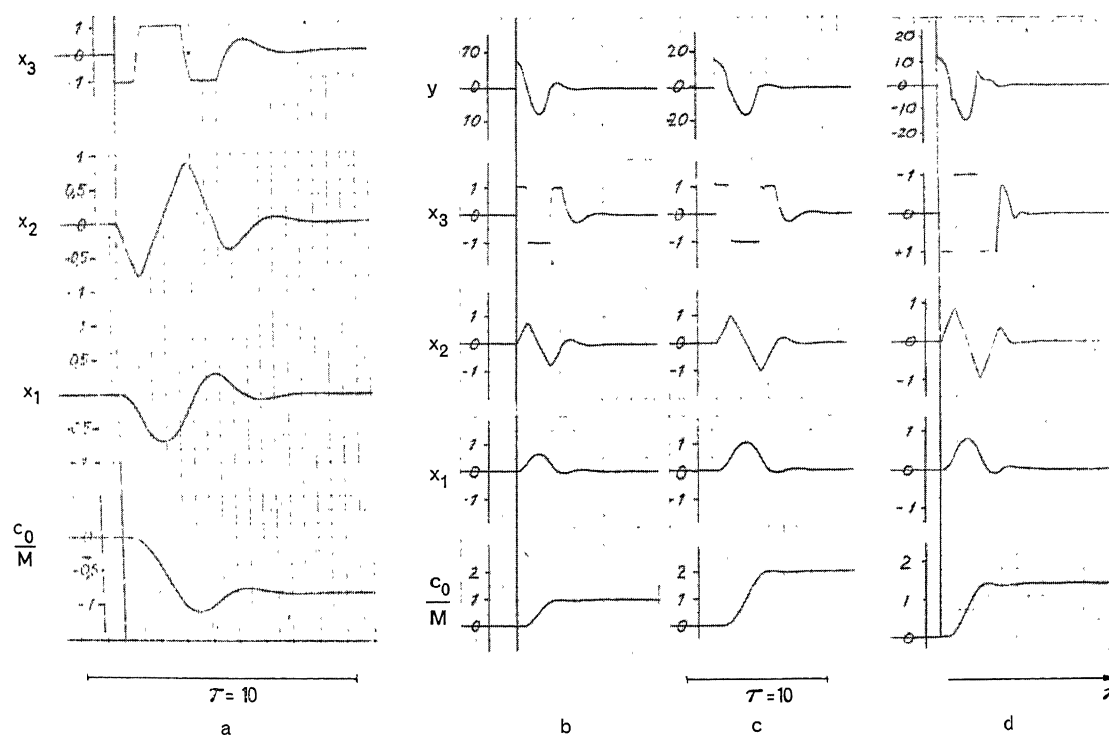


Figure 9. Step response for system with  $n = 3$ : (a) Linear controller; (b) non-linear controller; (c) non-linear controller; (d) non-linear controller, but on a system of higher order than third

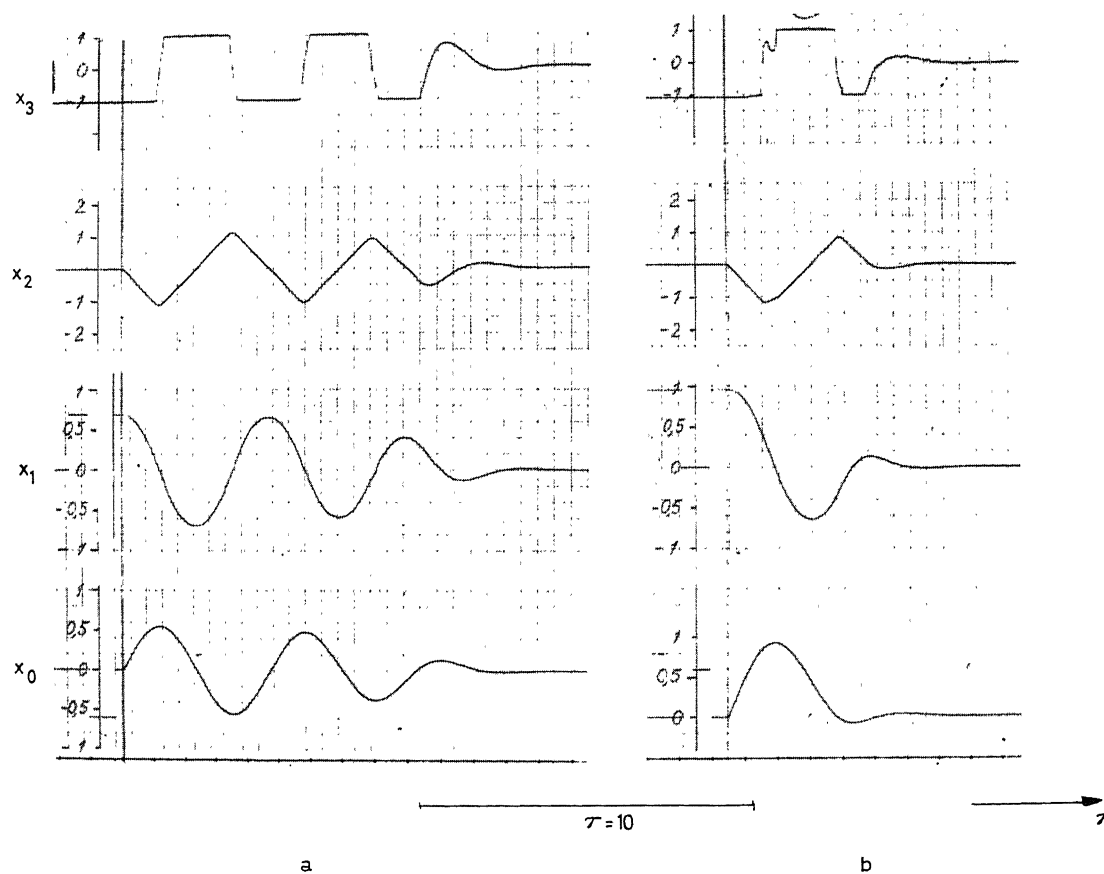


Figure 10. Ramp response for system with  $n = 3$ : (a) Linear controller; (b) non-linear controller

In the linear range:

$$\begin{aligned} a_0 &= 8 & f_0(x_0) &= 8x_0 \\ a_1 &= 8 & f_1(x_1) &= 8x_1 \\ a_2 &= 4 & f_2(x_2) &= 4x_2 \end{aligned}$$

In the non-linear range:

$$\begin{aligned} B &= 3 \\ k_0 &= a_0 = 8 & f_0(x_0) &= 8x_0 \\ k_1 &= 16\sqrt{2} \text{ [from eqn (25)]} & f_1(x_1) &= 16\sqrt{2} \frac{x_1}{|x_1|} |x_1|^{3/2} \\ k_2 &= \frac{20}{3} \text{ [from eqn (25)]} & f_2(x_2) &= \frac{20}{3} x_2^3 \end{aligned}$$

Figure 9 (a) shows the step response with a linear controller for such a large step that the stability has started to deteriorate. For an insignificantly increased step the system is quite unstable. Responses with the non-linear controller are shown in Figure 9 (b) and (c) for two steps, both larger than the largest step that the linear controller can handle. As is apparent these responses are almost ideal bang-bang responses, which is not surprising, since the system has been designed for bang-bang step response. Figure 10 (b) shows, however, that the system operates practically bang-bang even for other inputs. In this case,  $x_1$  has been given an initial value, approximately equal to 1 (corresponding to a ramp input). Under the same conditions a system with a linear controller is unstable. The response with a linear controller and an initial value for  $x_1$  immediately below the stability limit is shown in Figure 10 (a).

Case  $n = 4$

In the case of four integrators, it became apparent that the design on the basis of eqns (26) does not yield a satisfactory response. The system is unstable for  $B = 2$ . For  $B = 4$ , limit-cycle oscillations were experienced. The same applies for  $B = 6$ , although the amplitudes of the oscillations were smaller than for  $B = 4$ . It is possible that a further increase in  $B$  would have made the system stable, but the investigation was directed instead towards finding a combination of constants  $k_v$  giving acceptable response for  $B = 6$ .

The following feedback functions were selected:

In the linear range:

$$\begin{aligned} a_0 &= 4 & f_0(x_0) &= 4x_0 \\ a_1 &= 8 & f_1(x_1) &= 8x_1 \\ a_2 &= 8 & f_2(x_2) &= 8x_2 \\ a_3 &= 4 & f_3(x_3) &= 4x_3 \end{aligned}$$

In the non-linear range:

$$\begin{aligned} k_0 &= 4.8 & f_0(x_0) &= 4.8 \frac{x_0}{|x_0|} |x_0|^{3/2} \\ k_1 &= 30.5 & f_1(x_1) &= 30.5 \frac{x_1}{|x_1|} |x_1|^2 \\ k_2 &= 440 & f_2(x_2) &= 440 x_2^3 \\ k_3 &= 64 & f_3(x_3) &= 64 \frac{x_3}{|x_3|} |x_3|^6 \end{aligned}$$

These values of the constants  $k_v$  correspond to:

$$\frac{k_1}{k_0} = 6.35 \quad \frac{k_2}{k_0} = 91 \quad \frac{k_3}{k_0} = 13.3$$

whereas eqns (26) give for  $B = 6$ :

$$\frac{k_1}{k_0} = 14.5 \quad \frac{k_2}{k_0} = 53.5 \quad \frac{k_3}{k_0} = 6.3$$

Although eqns (26) do not provide  $k_v$  values that can be applied, they nevertheless yield values that do not differ from those selected more than by a factor of about 2. Equations (26) are therefore of value for selecting reasonably starting values when determining the constants  $k_v$  in an analogue computer.

A couple of step responses with a linear controller are shown in Figure 11 (a) and (b), the first one for a step that the system can just handle, the second one with a slightly increased step where the system becomes unstable.

Corresponding step responses with the non-linear controller are given in Figure 11 (c) and (d). Figure 11 (e) and (f) show a couple of step responses with a considerably larger amplitude. A comparison of Figure 11 (e) and (f) reveals that the responses have almost exactly the same form.

In Figure 11 (f) the ideal bang-bang response has been plotted as the dotted line in the recording of the controlled output  $c_0/M$ . Even if the system investigated for large amplitudes yields six switchings, while in an ideal bang-bang system only four switchings occur, the settling time is only about 50 per cent longer than the shortest possible.

No investigations have been made on systems with limitations in a higher derivative than the fourth one. This does not mean, however, that the application is limited to fourth order systems. On the contrary, applications have been made on higher order systems with additional linear transfer functions introduced immediately ahead of the non-linear function representing the limitation (at  $e$  in Figure 1). For example with three integrators behind the limitation function ( $n = 3$ ), practically ideal bang-bang response is obtained even in this case, provided that the linear system is synthesized according to Kessler<sup>3</sup> and that a small correction is made to the constants  $k_v$ . (This correction was obtained simply by assuming the limitation level  $M$  to be smaller than the real one by a factor of 1.4.) A step response for such a system is shown in Figure 9 (d). Thus the proposed type of controller can be used for a system of arbitrary order, as long as there are no limitations in derivatives of higher order than the fourth, and the transfer functions between the limitation and the controlled output are integrators.

## Conclusions

The investigation reported in this paper shows that it is possible to design a non-linear controller, operating linearly for small signals and also having a satisfactory performance for large signals when used on a process, comprising a number of integrators, and in which one of the derivatives of the controlled output is limited. The controller is simple, since the only difference in comparison with a conventional linear controller is that the feedbacks already existing in the latter are made non-linear, which can easily be achieved by means of biased diodes or non-linear resistors. The transition between linear and non-linear operation takes place smoothly without any need for switching devices.

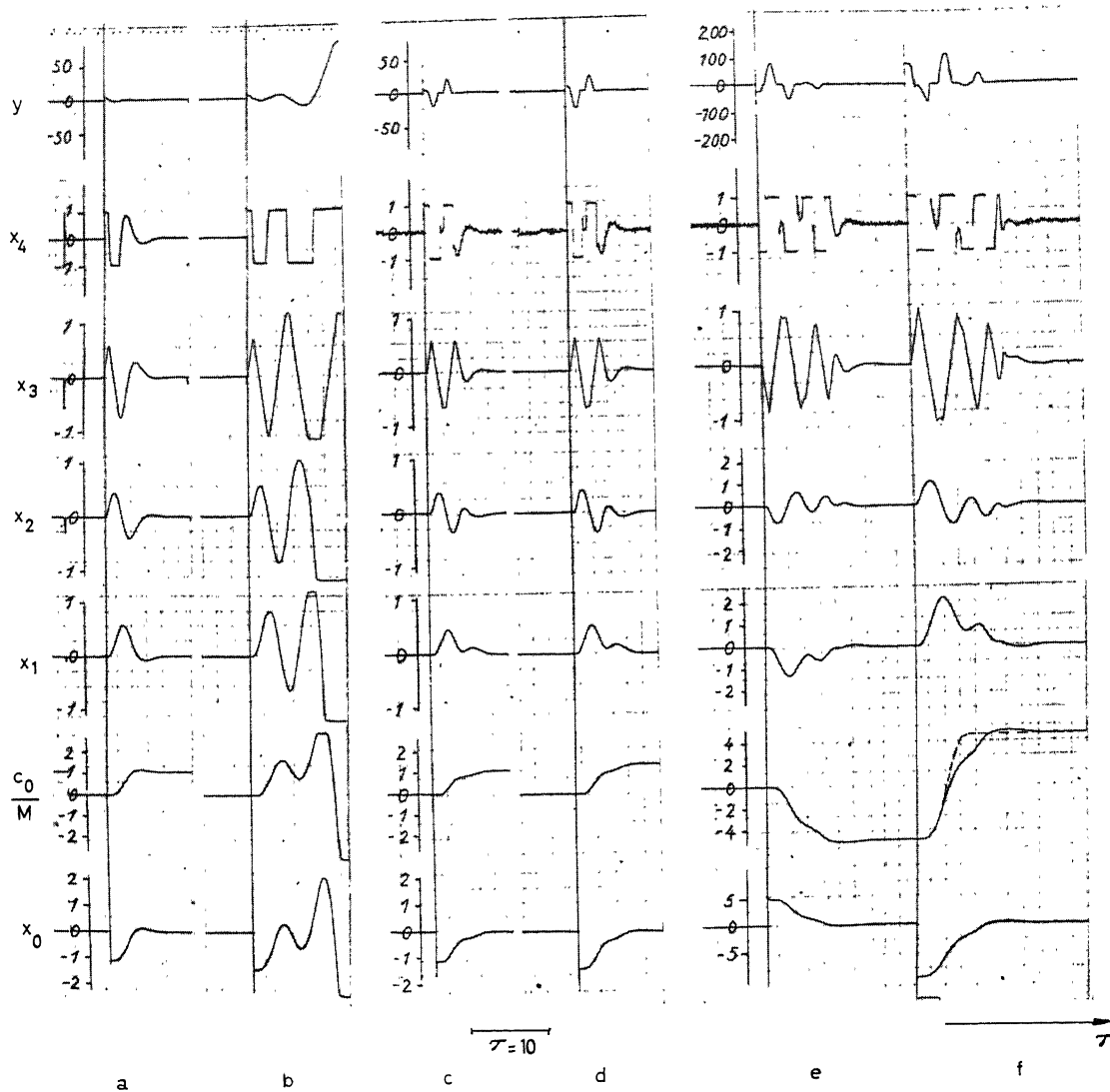


Figure 11. Step response for system with  $n = 4$ : (a) Linear controller; (b) linear controller; (c) non-linear controller; (d) non-linear controller; (e) non-linear controller; (f) non-linear controller; Dotted curve: ideal bang-bang response

An almost exact bang-bang response can be obtained for systems with limitation in the second or third derivative of the controlled output. In the case where the fourth derivative is limited, the response is sufficiently good not to justify a change-over to the extremely complicated controller required to achieve exact bang-bang response. By introducing the non-linear controller, improvements can also be achieved in the linear range, since the controller can be designed here for the best small-signal response without regard to the large-signal response. Even though the investigation presented here is confined to systems comprising solely integrators in the control loop, the results may be applied to cases where the integrators are replaced by transfer functions with poles sufficiently close to the origin in the  $s$  plane. It should therefore be possible to apply controllers of this type to a large number of processes.

#### Remaining Work

The studies are being continued along the following lines:

- Further investigations with  $n=3$  and  $4$ , but with  $B=1$  and  $2$
- Investigation of systems with more than one limited derivative.
- Investigations with other types of transfer functions than integrators.
- Investigation of how far it is possible to approximate the non-linear functions without losing the main advantages.

The author expresses his thanks to Mr. Hj. Sørensen who performed the investigations on the analogue computer.

#### References

- CHANDAKET, P., and LEONDES, C. T. Synthesis of quasi-stationary optimum control system. Pt I. *AIEE Trans. Applications and Industry*, No. 58 (1962) 313-319
- BURMEISTER, H. L. Zeitoptimale Übergangsvorgänge mit beschränkter  $n$ -ter Ableitung. *Z. Messen-Steuern-Regeln* 4, 10 (1961) 407-409
- KESSLER, C. Ein Beitrag zur Theorie mehrschleifiger Regelungen. *Regelungstechnik* 8, 8 (1960) 261-266

## DISCUSSION

M. ATHANS, *M.I.T. Lincoln Laboratory, Lexington, Mass. U.S.A.*

It should be emphasized that the system of *Figure 3* is *not* a time-optimal one. Since the word 'bang-bang' now implies time-optimal control it is somewhat confusing in this paper. I would like to ask the following questions.

Is the system of *Figure 3* stable for *all* possible initial conditions? How is *B* chosen? How can it be proved that the system is stable? Has the author investigated the relation of the response time to the minimum time?

F. MESCH, *Institut für Regelungstechnik, T.H. Darmstadt, Germany*

At the end of the paper it is stated that the results obtained for the integrators might be extended to systems with poles sufficiently close to the origin of the *s* plane. I do not believe that this is true for complex poles. Furthermore, I feel it to be somewhat unrealistic to assume that the output signals of all the integrators can be measured in a practical system.

E. PAVLIK, *Siemens, Lassallestr. 9, Karlsruhe, Germany*

According to *Figure 3* of the paper, the non-linear controller feedback functions are taken from the outputs of the integrators of the controlled system. Normally, in technical control systems, these signals are not available. What means exist for approximated bang-bang control, when these feedback signals are not available?

M. HAMZA, *ETH Zurich, Switzerland*

I would like to know how far it is possible to approximate the non-linear functions without losing the main advantages of this method? At the end of this paper it was mentioned that such studies were being performed. I feel that this method will have many limitations for most systems in practice. From the previous discussion, I understand that studies were performed, using a system having complex poles. Would Mr. Persson please indicate how he would apply his method if the open-loop system transfer function is of the form

$$G(s) = \frac{K}{s(s^2 + 2\xi\omega_0 s + \omega_0^2)}$$

E. V. PERSSON, *in reply*

In reply to Mr. Athans' remark concerning the word 'bang-bang', I refer to the title and the introduction of my paper, from which I hope it is clear that I am aware of the fact that the system is not a true bang-bang system (except for the case  $n = 2$ ). In my opinion this is less important than the fact that the system can be made stable without having to compromise between small and large signal response.

Investigations have indicated that the system of *Figure 3* can be designed to be stable for all initial conditions, even if this has not been strictly proved. The system  $n = 3$  and  $n = 4$  were checked for stability over a large domain of initial conditions, as well as for harmonic response over a large frequency range.

The choice of *B* is a compromise between practical considerations and the desire to approach bang-bang response as closely as possible. In my opinion  $B = 1$  or  $2$  should be preferred, even if the ratio between the response time and the minimum time will increase for a fourth-order system from 1.5, as obtained in *Figure 11* in my paper, to approximately 2.

Mr. Mesch and Dr. Hamza have drawn attention to systems with complex poles in the *s* plane. Systematic studies of such systems have not been carried out, but the method has been successfully applied to a fourth-order practical system of this kind. Briefly, the method was applied by selecting the exponents in the non-linear functions from the analogue computer diagram simulating the system, neglecting the feedbacks existing between the integrators. Then the constants  $k_i$  were determined by means of computer studies.

Mr. Mesch and Dr. Pavlik remarked that normally all the quantities necessary for feedback are not available in practical systems. This is true even with a linear controller. To be specific, let us assume a process, consisting of three integrators, where no derivatives of the output are available. In order to make the linear system stable, one then has to produce the first and second derivative by means of lead networks. There are no additional difficulties in applying the method proposed in my paper. In fact, the method has been applied to a practical system of this kind.

Finally, replying to Dr. Hamza's first question, the non-linear functions can be approximated by three straight lines, without losing the main advantages.

# Dual Input Systems with a Saturation Constraint

R. S. GAYLORD

## Summary

This paper extends the theory of dual-input systems to the problem of limiting the activity of an intermediate variable in the system in order to avoid saturation or excessive power consumption. The derivation of the general dual-input Wiener filter which includes the constraint is first outlined and discussed and then three degenerate cases are completely solved. The Lagrange multiplier approach is used to handle the constraint mathematically. To demonstrate completely the concepts involved, a numerical example for a case is shown in detail. The numerical example demonstrates how one might 'trade off' or compromise in an actual system design in order to find the 'best' solution.

## Sommaire

Ce travail étend la théorie de systèmes à entrée double au problème de limitation de l'activité d'une variable intermédiaire d'un système afin d'éviter la saturation ou la consommation excessive de puissance. La dérivation du filtre général de Wiener à entrée double, contenant la restriction est discutée d'abord; ensuite, des solutions complètes des trois cas dégénérés sont présentées. On se sert de la méthode du multiplicateur de Lagrange pour traiter le problème d'une manière mathématique. Pour présenter l'ensemble des concepts en question, un exemple numérique pour un cas particulier est donné. Cet exemple numérique montre comment on peut «échanger» ou arranger le calcul d'un système réel afin de trouver la solution «la meilleure».

## Zusammenfassung

Der Aufsatz erweitert die Theorie der Systeme mit zwei Eingängen auf folgendes Problem: In dem System soll die Auswirkung einer Zwischengröße begrenzt werden, um Sättigung oder übermäßige Leistungsaufnahme zu vermeiden. Die Darstellung eines Filters im Wiener'schen Sinne mit zwei Eingängen, das die Beschränkungen einschließt, wird zuerst abgeleitet sowie diskutiert und dann 3 (entartete) Fälle vollständig gelöst. Zur Lösung wird der Lagrange-Multiplikator verwendet, um die Beschränkungen (Nebenbedingungen) mathematisch zu berücksichtigen. Ein in allen Einzelheiten durchgerechnetes Beispiel erläutert eingehend das Verfahren. Das Zahlenbeispiel zeigt, wie man eine Abänderung oder einen Kompromiß beim Entwurf eines wirklichen Systems treffen muß, um zur besten Lösung zu gelangen.

## Introduction

The problem of extracting the best estimate of a distorted signal by using more than one source of information is of considerable engineering interest and has been examined by various authors in the past few years. The approach to the problem has generally taken two paths, the first being a direct derivation of optimum filters by using the criterion of minimum mean-squared error<sup>1</sup> and the other, the derivation of the so-called distortionless type of filter which eliminates the signal distortion component of error completely<sup>2, 3</sup>. This approach simplifies the mathematics and the design problem but frequently at the expense of a greater overall mean squared error at the output.

This paper extends dual input filter theory to the case of limiting an input into fixed elements in the system to avoid

saturation. The introduction of this constraint is handled mathematically by use of the Lagrange multiplier technique. The analysis considers an arbitrary fixed plant dynamics in order that it may be applied to the practical problem of control system design as well as the usual filtering and estimating problem.

The technique of combining 'redundant' measurements in order to improve the quality of estimation or control has not been used as widely as it could be in industry to date. This method will receive greater emphasis in the design of guidance and control systems for future missiles and space vehicles because component and other equipment weights are constantly being improved while accuracy and reliability are receiving greater emphasis. The appropriate use of redundancy is a technique which can do much to improve the accuracy and reliability of a system. Examples of future systems which could use independent measurements of related quantities are:

- (1) Satellite rendezvous terminal guidance system using independent measurements of range and range rate.
- (2) Ballistic missile or space booster guidance using a doppler-aided inertial guidance system (independent measurements of velocity).
- (3) Satellite attitude control system which would make use of independent measurements of attitude, such as horizon sensors, star trackers, sun sensors, magnetometers, gyros, etc.

These are only a few of the possible navigation and control applications of this technique. Space systems have been emphasized because of their extreme requirements for accuracy and reliability. Many other possible applications exist, however, where these techniques could be used profitably.

It is believed that this paper, in extending the work on the problem of design using multiple independent measurements to the very practical case of limiting to avoid saturation, increases the usefulness of the theory. Such a constraint on the mean squared value of a variable within the system is a way of restricting the activity of a key intermediate variable and hence, may be important in designs for space as a technique to conserve power or energy which are premium quantities in that environment.

## Description of the Problem

The general single-input problem of filter design with limiting to avoid saturation is discussed thoroughly in Chapter 7 of Ref. 4. This desire to avoid non-linear regions of operation in fixed elements in a system is an important constraint on control system designers. Another motivation for limiting the mean squared activity of a variable would be to limit the power consumed or other 'costs' which occur in the act of controlling a dynamical system. This paper extends the single-input theory with non-linear elements to dual input systems. The linear dual input theory is thus extended to non-linear systems through the device of constraining the mean squared activity of an intermediate variable.

The system to be considered is shown in *Figure 1*. The signals are stochastic variables and the constraint on the input to the fixed elements of the system,  $x(t)$ , will involve limiting the R.M.S. amplitude of the saturating signal,  $\bar{x}^2(t)$  to some specified fraction of the linear range of the saturating element  $G_f$ .

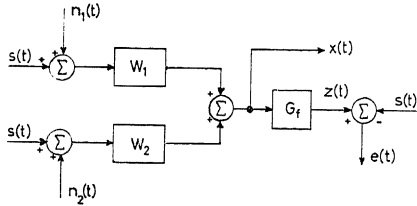
The problem of control system design with specified criteria is of considerable consequence<sup>5</sup>. The problem of minimizing the mean squared error is considered subject to the constraint

$$\bar{x}^2(t) \leq \sigma_s^2 \quad (1)$$

In order to select the form of compensation,  $W_1$  and  $W_2$ , that will minimize the mean squared error and meet the constraining relation, a functional may be formed by means of the Lagrange multiplier technique<sup>4, 5</sup>. This functional is

$$F = \bar{e}^2(t) + \lambda \bar{x}^2(t) \quad (2)$$

where  $\lambda$  is the Lagrangian multiplier. Since  $\bar{e}^2(t)$  and  $\bar{x}^2(t)$  may be written in terms of functionals of the unknowns  $W_1(\omega)$  and  $W_2(\omega)$ , application of the calculus of variations to the minimization of  $F$  yields a set of two simultaneous equations just as in the case with the linear problem<sup>1, 2</sup>. These simultaneous equations



*Figure 1. Block diagram of semi-free configuration dual input system with saturating variable  $x(t)$*

may be solved for  $W_1$  and  $W_2$  in terms of the angular frequency  $\omega$  and the Lagrange multiplier  $\lambda$ . The solution for  $W_1(\omega, \lambda)$  and  $W_2(\omega, \lambda)$  may be substituted into the expression for  $\bar{x}^2(t)$  and  $\lambda$  adjusted so that the constraint eqn (1) is satisfied. The value of  $\lambda$  thus found may be used to find the appropriate compensation  $W_1(\omega)$  and  $W_2(\omega)$ , and the value of  $\bar{e}^2(t)$ . These steps will now be developed in detail.

Examine *Figure 1*, the transformed equations governing the dynamical behaviour of the system may be written by inspection, where the symbol  $\omega$  is the transform complex variable.

$$X(\omega) = [S(\omega) + N_1(\omega)] W_1(\omega) + [S(\omega) + N_2(\omega)] W_2(\omega) \quad (3)$$

$$Z(\omega) = X(\omega) G_f(\omega) \quad (4)$$

$$E(\omega) = Z(\omega) - S(\omega) \quad (5)$$

The error may be written

$$E(\omega) = G_f(\omega) \{ [S(\omega) + N_1(\omega)] W_1(\omega) + [S(\omega) + N_2(\omega)] W_2(\omega) \} - S(\omega) \quad (6)$$

Making the following substitutions:

$$G_f(\omega) W_1(\omega) = Y_1(\omega) \quad (7)$$

$$G_f(\omega) W_2(\omega) = Y_2(\omega) \quad (8)$$

Eqn (6) becomes (dropping the functional notation)

$$E = Y_1 N_1 + Y_2 N_2 + [Y_1 + Y_2 - 1] S \quad (9)$$

Eqn (9) is in a familiar form occurring in linear filtering and prediction. It leads directly to the spectral density of the error<sup>6</sup>. If  $s(t)$ ,  $n_1(t)$  and  $n_2(t)$  are independent stationary random time series the spectral density of the error may be obtained in terms of  $Y_1$ ,  $Y_2$ , their complex conjugates denoted by an  $A^*$ ; and the power spectral densities of the noise and signal quantities denoted by  $\Phi_{N_1}$ ,  $\Phi_{N_2}$  and  $\Phi_S$ , respectively.

$$\Phi_E = |Y_1|^2 (\Phi_S + \Phi_{N_1}) + |Y_2|^2 (\Phi_S + \Phi_{N_2}) + \Phi_S (Y_1 Y_2^* + Y_1^* Y_2) - \Phi_S (Y_1^* + Y_2^* + Y_1 + Y_2) + \Phi_S \quad (10)$$

The mean squared value of the error is given by

$$\bar{e}^2 = \frac{1}{2\pi} \int_{-\infty}^{\infty} \Phi_E(\omega) d\omega \quad (11)$$

Thus, by substituting eqn (10) into eqn (11) the following may be obtained

$$\bar{e}^2 = \frac{1}{2\pi} \int_{-\infty}^{\infty} \{ [\Phi_S + \Phi_{N_1}] Y_1 Y_1^* + [\Phi_S + \Phi_{N_2}] Y_2 Y_2^* + \Phi_S [Y_1 Y_2^* + Y_1^* Y_2] - \Phi_S [Y_1^* + Y_2^* + Y_1 + Y_2] + \Phi_S \} d\omega \quad (12)$$

Similarly,

$$X = W_1 (S + N_1) + W_2 (S + N_2) \quad (13)$$

and

$$\bar{x}^2 = \frac{1}{2\pi} \int_{-\infty}^{\infty} [W_1 W_1^* (\Phi_S + \Phi_{N_1}) + W_2 W_2^* (\Phi_S + \Phi_{N_2}) + \Phi_S (W_1^* W_2 + W_1 W_2^*)] d\omega \quad (14)$$

Now the functional  $F$  may be formed by substitution of eqns (12) and (14) into eqn (2).

$$F = \frac{1}{2\pi} \int_{-\infty}^{\infty} [(\Phi_S + \Phi_{N_1}) |G_f|^2 W_1 W_1^* + (\Phi_S + \Phi_{N_2}) |G_f|^2 W_2 W_2^* + \Phi_S |G_f|^2 (W_1 W_2^* + W_1^* W_2) - \Phi_S G_f^* (W_1^* + W_2^*) - \Phi_S G_f (W_1 + W_2) + \Phi_S] d\omega + \frac{\lambda}{2\pi} \int_{-\infty}^{\infty} [(\Phi_S + \Phi_{N_1}) W_1 W_1^* + (\Phi_S + \Phi_{N_2}) W_2 W_2^* + \Phi_S (W_1^* W_2 + W_1 W_2^*)] d\omega \quad (15)$$

The calculus of variations may be used with this functional in order to find a minimum. Following the usual procedures of the calculus of variations to minimize  $F$ , allow the variation in  $W_1(\omega)$  to be  $\varepsilon \eta_1(\omega)$  and the variation in  $W_2(\omega)$  to be  $\varepsilon \eta_2(\omega)$  where  $\varepsilon$  is a parameter not depending upon  $\omega$  and  $\eta_1, \eta_2$  are arbitrary functions of  $\omega$ . At the same time  $F$  is subject to a variation  $\delta F$  and becomes  $F + \delta F$ . A necessary condition for the quantity  $F$  to have an extremum is

$$\frac{\partial}{\partial \varepsilon} [F + \delta F]_{\varepsilon \rightarrow 0} = 0 \quad (16)$$

for arbitrary  $\eta_1(\omega)$  and  $\eta_2(\omega)$ .

Replacing  $W_1(\omega)$  and  $W_2(\omega)$  with their respective variations and carrying out the operation indicated in eqn (16) it is found

$$W_1 |G_f|^2 (\Phi_S + \Phi_{N_1}) + W_2 |G_f|^2 \Phi_S - G_f^* \Phi_S + \lambda W_1 (\Phi_S + \Phi_{N_1}) + \lambda W_2 \Phi_S = Q_1 \quad (17)$$

$$W_2 |G_f|^2 (\Phi_S + \Phi_{N_2}) + W_2 |G_f|^2 \Phi_S - G_f^* \Phi_S + \lambda W_2 (\Phi_S + \Phi_{N_2}) + \lambda W_1 \Phi_S = Q_2 \quad (18)$$

The formal steps are given in detail elsewhere<sup>5,7</sup>. The result obtained depends upon the fact that the resulting integrals must be equal to zero if a minimum is to be found; and in order for this to be the case, each integral must contain products of functions whose singularities are in the same half plane. Thus, the quantities on the left-hand side of eqns (17) and (18) must contain no upper half plane poles. It is observed, therefore, that  $Q_1(\omega)$  and  $Q_2(\omega)$  are functions with lower half plane poles only.

Eqns (17) and (18) may be written

$$(\Phi_S + \Phi_{N_1}) W_1 (\lambda + |G_f|^2) + \Phi_S W_2 (\lambda + |G_f|^2) = Q_1 + G_f^* \Phi_S \quad (19)$$

$$(\Phi_S + \Phi_{N_2}) W_2 (\lambda + |G_f|^2) + \Phi_S W_1 (\lambda + |G_f|^2) = Q_2 + G_f^* \Phi_S \quad (20)$$

To proceed, spectral factorization<sup>4</sup> must be applied to the factor  $(\lambda + |G_f|^2)$ . (In actual practice this is not always easy to do, it is useful to note that common fixed plants which are readily subject to this factorization are of the form  $K/S^n$  or a simple lag.)

$$(\lambda + |G_f|^2) = \alpha^+ \alpha^- \quad (21)$$

Thus

$$(\Phi_S + \Phi_{N_1}) W_1 \alpha^+ + \Phi_S W_2 \alpha^+ = \frac{Q_1}{\alpha^-} + \frac{G_f^* \Phi_S}{\alpha^-} \quad (22)$$

$$(\Phi_S + \Phi_{N_2}) W_2 \alpha^+ + \Phi_S W_1 \alpha^+ = \frac{Q_2}{\alpha^-} + \frac{G_f^* \Phi_S}{\alpha^-} \quad (23)$$

These equations now reduce to the following simple form:

$$(\Phi_S + \Phi_{N_1}) R_1 + \Phi_S R_2 = Q_1' + \phi_0 \quad (24)$$

$$(\Phi_S + \Phi_{N_2}) R_2 + \Phi_S R_1 = Q_2' + \phi_0 \quad (25)$$

where

$$Q_1' = Q_1 / \alpha^- \quad (26)$$

$$Q_2' = Q_2 / \alpha^- \quad (27)$$

and

$$R_1 = W_1 \alpha^+ \quad (28)$$

$$R_2 = W_2 \alpha^+ \quad (29)$$

$$\phi_0 = \frac{G_f^* \Phi_S}{\alpha^-} \quad (30)$$

### Degenerate Solutions

To apply the method of undefined coefficients at this point as was done by Wescott<sup>1</sup> would be extremely laborious, if not impossible, as the Lagrange multiplier is also unknown. In many practical applications of this theory one may be perfectly free to assume similarity of bandwidth either for signal and noise or for the two noise inputs. By selecting the similar power spectral densities to be identical in bandwidth a more conservative approach may be made and the problem becomes tractable once again. This technique is common in engineering practice where much of the quantitative data used in a design problem

corresponds to conservative estimates obtained from experimentation or theory. One can proceed by examining the so-called 'degenerate cases' of Stewart and Parks<sup>2</sup>.

### Case I

First, consider the case where

$$\Phi_{N_2} = K^2 \Phi_{N_1} \quad (31)$$

Substituting this into eqn (25)

$$(\Phi_S + K^2 \Phi_{N_1}) R_2 + \Phi_S R_1 = Q_2' + \phi_0 \quad (32)$$

Subtracting eqn (28) from (20)

$$R_1 \Phi_{N_1} - R_2 K^2 \Phi_{N_1} = Q_1' - Q_2' \quad (33)$$

and so

$$\Phi_{N_1} (R_1 - K^2 R_2) = Q_3 \quad (34)$$

where  $Q_3$  is  $Q_1' - Q_2'$  and has poles only in the lower half plane.

If  
then

$$\Phi_{N_1} = \phi_1^+ \phi_1^- \quad (35)$$

$$\phi_1^+ (R_1 - K^2 R_2) = \frac{Q_3}{\phi_1^-} \quad (36)$$

Observe that the left side of eqn (36) can have poles only in the upper half plane and the right side can have poles only in the lower half plane, hence

$$\phi_1^+ (R_1 - K^2 R_2) = 0 \quad (37)$$

$$R_1 = K^2 R_2 \quad (38)$$

Substitution of eqn (38) into eqn (24) obtains

$$\left( \Phi_S + \frac{\Phi_{N_2}}{K^2} \right) K^2 R_2 + \Phi_S R_2 = Q_1' + \phi_0 \quad (39)$$

and factoring,

$$\Phi_S R_2 (1 + K^2) + R_2 \Phi_{N_2} = Q_1' + \phi_0 \quad (40)$$

Let

$$R_2' = (1 + K^2) R_2 \quad (41)$$

$$\Phi_{N_2}' = \frac{1}{1 + K^2} \Phi_{N_2} \quad (42)$$

then

$$R_2' (\Phi_S + \Phi_{N_2}') = Q_1' + \phi_0 \quad (43)$$

Next form

$$\Phi_S + \Phi_{N_2}' = \phi^+ \phi^- \quad (44)$$

thus

$$R_2' \phi^+ = \frac{Q_1'}{\phi^-} + \frac{\phi_0}{\phi^-} \quad (45)$$

Observe that the right side of eqn (45) contains upper half plane poles only and thus this equation leads directly to the Wiener-Kolmogoroff solution for the optimum filter<sup>5</sup>. Hence, after substituting for  $\phi_0$  the relation is obtained

$$R_2'(\omega, \lambda) = \frac{1}{\phi^+(\omega)} \left[ \frac{G_f^*(\omega) \Phi_S(\omega)}{\alpha^-(\omega, \lambda) \phi^-(\omega)} \right]_+ \quad (46)$$

The bracket symbol  $[ ]_+$  is taken to mean that a partial fraction expansion is formed and the terms with lower half plane roots are discarded. The filters which minimize  $\bar{e}^2$  subject to the constraint that  $\bar{x}^2 \leq \sigma_s^2$  may be found as a function of the Lagrange multiplier by substituting eqn (29) into eqn (46),

$$W_2(\omega, \lambda) = \frac{1}{1+K^2} \left[ \frac{1}{\alpha^+(\omega, \lambda) \phi^+(\omega)} \right] \left[ \frac{G_f^*(\omega) \Phi_S(\omega)}{\alpha^-(\omega, \lambda) \phi^-(\omega)} \right]_+ \quad (47)$$

From eqns (28), (29) and (38) it is found that

$$W_1 \alpha^+ = K^2 W_2 \alpha^+ \quad (48)$$

or

$$W_1 = K^2 W_2 \quad (49)$$

Next, substitute these filter functions into eqn (14) and find the value of  $\lambda$  which satisfies the constraint on  $x^2$ .

### Case II

The next problem is to derive the solution for the case where it may be assumed

$$\Phi_{N_2} = c^2 \Phi_S \quad (50)$$

With this substitution, eqns (24) and (25) become

$$(\Phi_S + \Phi_{N_1}) R_1 + \Phi_S R_2 = Q'_1 + \phi_0 \quad (51)$$

$$\Phi_S R_1 + (\Phi_S + c^2 \Phi_S) R_2 = Q'_2 + \phi_0 \quad (52)$$

Factoring eqn (52)

$$\Phi_S [R_1 + R_2 (1 + c^2)] = Q'_2 + \phi_0 \quad (53)$$

and letting

$$\Phi_S = \phi^+ \phi^-$$

find

$$\phi^+ [R_1 + R_2 (1 + c^2)] = \frac{Q'_2}{\phi^-} + \frac{\phi_0}{\phi^-} \quad (54)$$

The left-hand side of eqn (54) has poles in the upper half plane only. The first term of the right-hand side has poles in the lower half plane only, hence as before it must be dropped from the equation. From eqn (54) is obtained

$$R_2 = \left( \frac{1}{1+c^2} \right) \left( \frac{1}{\phi^+} \left[ \frac{\phi_0}{\phi^-} \right]_+ - R_1 \right) \quad (55)$$

Although this equation is somewhat complicated, the only unknown quantities now appearing are  $R_1$  and  $R_2$ , the unknown  $Q'_2$  having been eliminated. For notational convenience define

$$\beta(\omega, \lambda) = \left( \frac{1}{1+c^2} \right) \frac{1}{\phi^+} \left[ \frac{\phi_0}{\phi^-} \right]_+ \quad (56)$$

Now, substitute eqns (55) and (56) into eqn (51) and obtain

$$(\Phi_S + \Phi_{N_1}) R_1 + \Phi_S \left( \beta - \frac{R_1}{1+c^2} \right) = Q'_1 + \phi_0 \quad (57)$$

Factoring, obtain

$$\Phi_S R_1 \left( \frac{c^2}{1+c^2} \right) + \Phi_{N_1} R_1 = Q'_1 + \phi_0 - \Phi_S \beta \quad (58)$$

Next, define some new functions

$$R'_1 = \frac{c^2}{1+c^2} R_1 \quad (59)$$

$$\Phi'_{N_1} = \frac{1+c^2}{c^2} \Phi_{N_1} \quad (60)$$

Substituting eqns (59) and (60) into eqn (58) find

$$\Phi_S R'_1 + \Phi'_{N_1} R'_1 = Q'_1 + \phi_0 - \Phi_S \beta \quad (61)$$

Let

$$(\Phi_S + \Phi'_{N_1}) = \phi_1^+ \phi_1^- \quad (62)$$

then

$$\phi_1^+ R'_1 = \frac{Q'_1}{\phi_1^-} + \frac{\phi_0}{\phi_1^-} - \frac{\Phi_S \beta}{\phi_1^-} \quad (63)$$

The first term on the right-hand side of eqn (63) contains only lower half plane poles. The remainder of the equation may be made compatible with the left-hand side to obtain

$$R'_1 = \frac{1}{\phi^+} \left[ \frac{\phi_0}{\phi_1^-} - \frac{\Phi_S \beta}{\phi_1^-} \right]_+ \quad (64)$$

Applying eqns (28) and (59) the final solution for  $W_1(\omega, \lambda)$  may be obtained

$$W_1(\omega, \lambda) = \frac{1}{\alpha^+(\omega, \lambda)} \left[ \frac{1+c^2}{c^2} \right] \left[ \frac{1}{\phi^+(\omega)} \right] \left[ \frac{\phi_0(\omega, \lambda)}{\phi_1^-(\omega, \lambda)} - \frac{\Phi_S(\omega) \beta(\omega, \lambda)}{\phi_1^-(\omega, \lambda)} \right]_+ \quad (65)$$

Combining eqns (29) and (55) find

$$W_2 \alpha^+ = \beta - \frac{R_1}{1+c^2} \quad (66)$$

and hence

$$W_2 = \frac{\beta}{\alpha^+} - \frac{W_1}{1+c^2} \quad (67)$$

As before, these filter functions are substituted into eqn (14) and the value of  $\lambda$  which satisfies the constraint on  $\bar{x}^2$  is found. This value of  $\lambda$  then enables the evaluation of  $W_1$  and  $W_2$  and the  $\bar{e}^2$ .

### Case III: The Distortionless Filter

The techniques for the design of the distortionless filter with a saturation constraint will now be developed. This type of filter is somewhat simpler to analyse and design than the Wiener filter. There are many practical applications for this type of filter, particularly if the signal power spectral density is not known or if the signal is deterministic rather than stochastic in nature.

Examining Figure 1, it is observed that the error is given by eqn (9) where the symbols are defined as before. Eqn (9) is repeated here for convenience,

$$E = Y_1 N_1 + Y_2 N_2 + [Y_1 + Y_2 - 1] S \quad (9)$$

In order for this filter to be 'distortionless' the last term on the right-hand side of eqn (9) must be zero, hence

$$Y_1 + Y_2 = 1 \quad (68)$$



and now the error is due to noise terms only,

$$E = Y_1 N_1 + (1 - Y_1) N_2 \quad (69)$$

Eqn (69) permits the finding of the power density of the error.

$$\Phi_E(\omega) = |Y_1(\omega)|^2 \Phi_{N_1}(\omega) + |1 - Y_1(\omega)|^2 \Phi_{N_2}(\omega) \quad (70)$$

The functional of eqn (2) remains the same with the mean of  $\bar{e}^2$  changing to that of a distortionless filter. Thus,

$$F = \frac{1}{2\pi} \int_{-\infty}^{\infty} \Phi_E(\omega) d\omega + \frac{\lambda}{2\pi} \int_{-\infty}^{\infty} \Phi_X(\omega) d\omega \quad (71)$$

For this case then, the functional to which the calculus of variations must be applied may be written by substitution into eqn (71).

$$\begin{aligned} F = & \frac{1}{2\pi} \int_{-\infty}^{\infty} (|Y_1|^2 \Phi_{N_1} + |Y_2|^2 \Phi_{N_2}) d\omega \\ & + \frac{\lambda}{2\pi} \int_{-\infty}^{\infty} [|W_1|^2 (\Phi_S + \Phi_{N_1}) + |W_2|^2 (\Phi_S + \Phi_{N_2}) \\ & + W_1^* W_2 \Phi_S + W_1 W_2^* \Phi_S] d\omega \end{aligned} \quad (72)$$

Recalling that  $Y_1 = G_f W_1$  and  $Y_2 = G_f W_2$  apply the calculus of variations to find an extremum of eqn (72). This results in deriving the following two simultaneous equations:

$$W_1 |G_f|^2 \Phi_{N_1} + \lambda W_1 (\Phi_S + \Phi_{N_1}) + \lambda W_2 \Phi_S = -Q_1 \quad (73)$$

$$W_2 |G_f|^2 \Phi_{N_2} + \lambda W_2 (\Phi_S + \Phi_{N_2}) + \lambda W_1 \Phi_S = -Q_2 \quad (74)$$

Eqn (68) may be written

$$W_1 + W_2 = \frac{1}{G_f} \quad (75)$$

After algebraic manipulation and the use of eqn (75) obtain

$$\Phi_{N_1} W_1 (\lambda + |G_f|^2) + \Phi_S \lambda G_f^{-1} = -Q_1 \quad (76)$$

$$\Phi_{N_2} W_2 (\lambda + |G_f|^2) + \Phi_S \lambda G_f^{-1} = -Q_2 \quad (77)$$

Subtraction of eqn (77) from (76) obtains

$$(\lambda + |G_f|^2) (\Phi_{N_1} W_1 - \Phi_{N_2} W_2) = Q_2 - Q_1 \quad (78)$$

Define some new functions

$$\Phi'_{N_1} = (\lambda + |G_f|^2) \Phi_{N_1} \quad (79)$$

$$\Phi'_{N_2} = (\lambda + |G_f|^2) \Phi_{N_2} \quad (80)$$

$$Q_3 = Q_2 - Q_1 \quad (81)$$

Eqn (78) now takes the form

$$\Phi'_{N_1} W_1 - \Phi'_{N_2} W_2 = Q_3 \quad (82)$$

Using eqn (75) find

$$\Phi'_{N_1} W_1 + \Phi'_{N_2} (W_1 - G_f^{-1}) = Q_3 \quad (83)$$

Hence

$$(\Phi'_{N_1} + \Phi'_{N_2}) W_1 = Q_3 + \Phi'_{N_2} G_f^{-1} \quad (84)$$

Letting

$$(\Phi'_{N_1} + \Phi'_{N_2}) = \phi^+ \phi^- \quad (85)$$

it is found

$$\phi^+ W_1 = \frac{Q_3}{\phi^-} + \frac{\Phi'_{N_2} G_f^{-1}}{\phi^-} \quad (86)$$

Notice that the situation again exists where the left side of eqn (86) contains only upper half plane poles and only a portion of the second term on the right may be retained. The portion of the second term is found by applying the methods of Wiener and Kolmogoroff and the resulting optimum filter transfer function is

$$W_1(\omega, \lambda) = \frac{1}{\phi^+} \left[ \frac{\Phi_{N_2} G_f^{-1}}{\phi^-} \right]_+ \quad (87)$$

Also

$$W_2(\omega, \lambda) = \frac{1}{G_f} - W_1(\omega, \lambda) \quad (88)$$

These transfer functions are substituted into eqn (14). Then the constraint relation eqn (1) is applied by selecting that value of  $\lambda$  which satisfies the constraint. This value of  $\lambda$  may now be used to evaluate  $W_1$  and  $W_2$  and then eqns (70) and (11) provide the value of  $\bar{e}^2$ . Notice that if this value of the mean squared error is too great for a reasonable system design, a compromise constraint value may be chosen. In fact, in a practical design problem one would undoubtedly examine the relation between the value of constraint on power (or saturation) as compared with the mean squared error resulting from such a constraint. The following numerical example should clarify the above concepts.

### Numerical Example

Examine Figure 2 which represents the example problem. Select  $G_f = 1.0$  (no dynamic portions). This will reduce the algebraic manipulations and thus aid in clarifying the work. For this example assume

$$\Phi_{S_2}(\omega) = \frac{4}{\omega^2 + 1} = \Phi_S(\omega)$$

$$\Phi'_{N_1}(\omega) = \frac{1}{\omega^2 + 1}; \quad \Phi_{N_1}(\omega) = \frac{1}{\omega^2(\omega^2 + 1)}$$

$$\Phi_{N_2}(\omega) = \frac{2}{\omega^2 + 1}$$

Following the techniques outlined in Case III (the distortionless filter) evaluate eqns (79), (80).

$$\Phi'_{N_1} = (\lambda + 1) \left[ \frac{1}{\omega^2(\omega^2 + 1)} \right] \quad (89)$$

$$\Phi'_{N_2} = (\lambda + 1) \left[ \frac{2}{\omega^2 + 1} \right] \quad (90)$$

and eqn (86) indicates

$$\begin{aligned} \phi^+ \phi^- &= (\Phi'_{N_1} + \Phi'_{N_2}) \\ &= (\lambda + 1) \left[ \frac{1 + \sqrt{2} j\omega}{j\omega(1 + j\omega)} \right] \left[ \frac{1 - \sqrt{2} j\omega}{-j\omega(1 - j\omega)} \right] \end{aligned} \quad (91)$$

Following eqn (87) find

$$\left[ \frac{\Phi'_{N_2} G_f^{-1}}{\phi^-} \right]_+ = \left[ \frac{2(\lambda + 1)(-j\omega)}{(1 + j\omega)(1 - \sqrt{2} j\omega)} \right]_+ \quad (92)$$

and expanding eqn (92) obtain

$$\frac{\Phi'_{N_2} G_f^{-1}}{\phi^-} = \frac{K_1}{1+j\omega} + \frac{K_2}{1-\sqrt{2}j\omega} \quad (93)$$

then

$$K_1 = \frac{2\lambda+2}{1+\sqrt{2}} = C_0(\lambda) \quad (94)$$

Now evaluating eqn (87)

$$W_1(\omega, \lambda) = C_0(\lambda) \left[ \frac{j\omega}{1+\sqrt{2}j\omega} \right] \quad (95)$$

and from eqn (88)

$$W_2(\omega, \lambda) = 1 - \frac{C_0(\lambda)j\omega}{1+\sqrt{2}j\omega} = \frac{1+(\sqrt{2}-C_0)j\omega}{1+\sqrt{2}j\omega} \quad (96)$$

From eqn (1)

$$\overline{x^2}(t) \leq \sigma_s^2 \geq \frac{1}{2\pi j} \int_{-j\infty}^{+j\infty} \Phi_X(s) ds \quad (97)$$

where  $s$  is the Laplace complex variable. Next, write eqn (97) with reference to eqn (14),

$$\begin{aligned} \overline{x^2}(t) = & \frac{1}{2\pi j} \int_{-j\infty}^{+j\infty} \{W_1(s)W_1(-s)[\Phi_S(s) + \Phi_{N_1}(s)] \\ & + W_2(s)W_2(-s)[\Phi_S(s) + \Phi_{N_2}(s)] \\ & + W_1(s)W_2(-s)\Phi_S(s) + W_1(-s)W_2(s)\Phi_S(s)\} ds \end{aligned} \quad (98)$$

Breaking the integral into three parts,

$$\left. \begin{aligned} J_1 &= \frac{1}{2\pi j} \int_{-j\infty}^{+j\infty} |W_1|^2 (\Phi_S + \Phi_{N_1}) ds \\ J_2 &= \frac{1}{2\pi j} \int_{-j\infty}^{+j\infty} |W_2|^2 (\Phi_S + \Phi_{N_2}) ds \\ J_3 &= \frac{1}{2\pi j} \int_{-j\infty}^{+j\infty} (W_1 W_2^* \Phi_S + W_1^* W_2 \Phi_S) ds \end{aligned} \right\} \quad (99)$$

Substituting for  $W_1$  and  $W_2$  from eqns (95) and (96) the integral is evaluated to obtain

$$\left. \begin{aligned} J_1 &= \frac{C_0^2(\lambda)}{2\sqrt{2}} \\ J_2 &= \frac{3C_0^2(\lambda)}{9.656} + \frac{6\sqrt{2}C_0(\lambda)}{9.656} + 1.25 \\ J_3 &= \frac{\sqrt{2}C_0^2(\lambda)}{4.428} \end{aligned} \right\} \quad (100)$$

Summing these and applying eqn (97),  $C_0(\lambda)$  must satisfy the following relation

$$0.957 C_0^2 + 0.877 C_0 + 1.25 \leq \sigma_s^2 \quad (101)$$

From eqn (101) it is obvious that [see eqn (96) and invoke physical realizability]

$$4(0.957)(1.25 - \sigma_s^2) \leq (0.877)^2 \quad (102)$$

Therefore,

$$\sigma_s^2 \geq 1.05 \quad (103)$$

in order for a solution to be possible at all. This corresponds to the lowest value of constraint which can be met. This solution corresponds to the following filter functions,

$$W_1(\omega) = \frac{-0.453 j\omega}{1 + \sqrt{2} j\omega} \quad (104)$$

$$W_2(\omega) = \frac{1 + 1.867 j\omega}{1 + \sqrt{2} j\omega} \quad (105)$$

This solution would be mechanized by Bendat's method<sup>3</sup> as shown in Figure 3. By the more direct method (after Figure 2) it would take the form of Figure 4. In these figures  $p$  is meant to represent  $d/dt$  and replaces  $j\omega$ . Eliminating the integrator and recognizing the fact that in Figure 2 an integration was assumed in the channel containing  $W_1(p)$ , for mathematical convenience, it is possible to show another mechanization for the filter which is that of Figure 4.

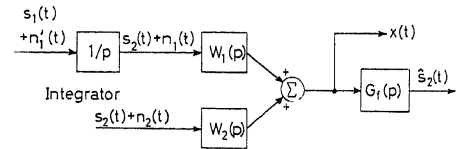


Figure 2. Block diagram of example problem

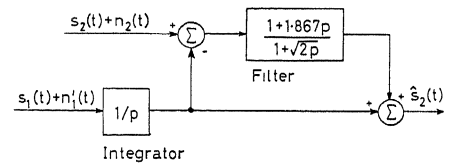


Figure 3. A possible mechanization of the example

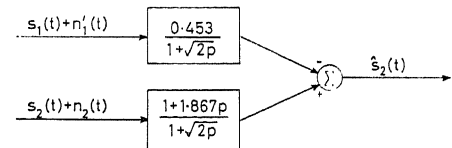


Figure 4. Another mechanization of the example

Next, examine how the constraint has affected the mean squared error. For this filter write  $\bar{e}^2$  from eqns (70) and (11)

$$\bar{e}^2 = \frac{1}{2\pi j} \int_{-j\infty}^{+j\infty} \{|W_1(s)|^2 \Phi_{N_1}(s) + |1 - W_1(s)|^2 \Phi_{N_2}(s)\} ds \quad (106)$$

and evaluating this integral we obtain

$$\bar{e}^2 = 1.48 \quad (107)$$

The minimum mean squared error may be found from eqn (72) with  $\lambda$  equal to zero. The minimization procedure for the distortionless filter, when it is not constrained to avoid saturation, is straightforward<sup>7</sup> and is therefore not shown here. The minimum mean squared error for the unconstrained case is

$$\bar{e}^2|_{\min} = 0.670 \quad (108)$$

This is a value less than half of that given in eqn (107).

The activity to be expected of the constrained variable  $x$  may be found for the unconstrained case by using the fact that

$$\overline{x^2} = \overline{e^2} + \frac{1}{2\pi j} \int_{-j\infty}^{+j\infty} \Phi_S(s) ds \quad (109)$$

Evaluating the integral and adding eqn (108)

$$\overline{x^2} = 2.67 \quad (110)$$

Thus the bound on the constraint  $\sigma_s^2$  is such that all values of  $\sigma_s^2$  greater than 2.67 may be satisfied by the 'optimum' distortionless filter.

The author gratefully acknowledges the helpful advice on this problem received from Professor C. T. Leondes, Department of Engineering, University of California, Los Angeles, California.

## References

- 1 WESTCOTT, J. H. Design of multivariable optimum filters. *Trans. Amer. Soc. mech. Engrs* 80 (1958) 463
- 2 STEWART, R. M. and PARKS R. J. Degenerate solutions and an algebraic approach to the multiple-input linear filter design problem. *Trans. Inst. Radio Engrs, N.Y.* CT-4, (1957) 10
- 3 BENDAT, J. S. Optimum filters for independent measurement of two related perturbed messages. *Trans. Inst. Radio Engrs, N. Y., CT-4* (1957) 14
- 4 NEWTON, G. C., Jr., GOULD, L. A. and KAISER, J. F. *Analytical Design of Linear Feedback Controls*. 1957. New York; Wiley
- 5 TSIEN, H. S. *Engineering Cybernetics*. 1954. McGraw-Hill
- 6 LANING, J. H. and BATTIN, R. H. *Random Processes in Automatic Control*. 1956. McGraw-Hill
- 7 GAYLORD, R. S. The smoothing of two related perturbed messages with constraints. *M.S. Thesis*, University of California, Los Angeles. December 1960

## DISCUSSION

D. ŠILJAK, *Electrotechnical Faculty, University of Belgrade, Mačvanska 8, Belgrade, Yugoslavia*

This paper is a useful contribution to the literature in that it extends the optimization of mean squared error with saturation constraint<sup>1</sup> to the design of dual-input filters. The proposed design procedure can be applied to automatic control systems with multiple independent measurements which have certain important advantages. However, in the application of the procedure to the design of control systems, problems may arise due to the fact that the stability aspect of the control problem is neglected. To hold the peak value of a signal within the linear range by controlling its R.M.S. value may be inadequate and the control system would exhibit poorly damped transient modes. Moreover, it is often difficult to determine the appropriate value of the R.M.S. constraint; even the controlling of the maximum value of the R.M.S. signal effectively limits the tendency of the signal to exceed its linear range.

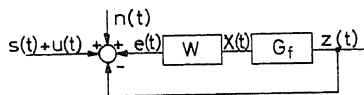


Figure A

By application of the generalized Mitrović method<sup>2</sup>, the problem of minimizing the mean squared error may be considered as subject to the constraint that the degree of relative stability is better, or at least equal to the prescribed value<sup>3</sup>. The minimization procedure will be shown briefly by the following example.

Consider the control system shown in Figure A, with the specifications

$$G_f(s) = \frac{1}{s(0.001s^2 + 0.025s + 0.25)} \quad (1)$$

$$W(s) = \frac{K(s + \lambda\delta)}{s + \delta}; \quad \phi_S(s) = -\frac{\gamma_S}{\pi s^2}$$

$$u(t) = \text{unit step function}; \quad \phi_N(s) = \gamma_N/\pi$$

It is required to determine the parameters  $K$ ,  $\lambda$  and  $\delta$  of the integral compensator  $W(s)$  so as to increase the system velocity constant and minimize the mean squared error while maintaining the overshoot of the step-function response below 30 per cent of its steady-state value. At first the noise component is not considered.

As known, the choice of the parameter  $\delta$  ( $\ll 1$ ) is not critical and the value 0.04 may be accepted. Then, if the numerical value of  $\gamma_S$  is  $2\pi$ , the power density of error  $\phi_E(s)$  corresponding to the signal  $s(t)$  is given as

$$\phi_E(s) = \frac{0.002s^3 + 0.05s^2 + 0.5s + 0.02}{0.001s^4 + 0.025s^3 + 0.25s^2 + (K + 0.01)s + 0.04K\lambda} \quad (2)$$

The denominator of  $\phi_E(s)$  is the characteristic polynomial of the system under investigation. By substituting  $K + 0.01 = \xi$  and  $0.04K = \eta$  in eqn (2), one may plot in the usual fashion<sup>2-4</sup> the characteristic curves  $\Gamma\zeta$  as shown in Figure B. These curves determine in the  $\phi\xi\eta$  plane the relative damping region which corresponds to certain value of the relative damping coefficient  $\zeta$ . So, for  $\zeta = 0.4$  the relative damping region is determined by the curve  $\Gamma_{0.4}$  and shown shaded in Figure B. For the values of  $\xi$  and  $\eta$  lying in this region all roots of the corresponding characteristic equation will have the relative damping coefficient greater than 0.4.

Applying the above substitution and using the well-known

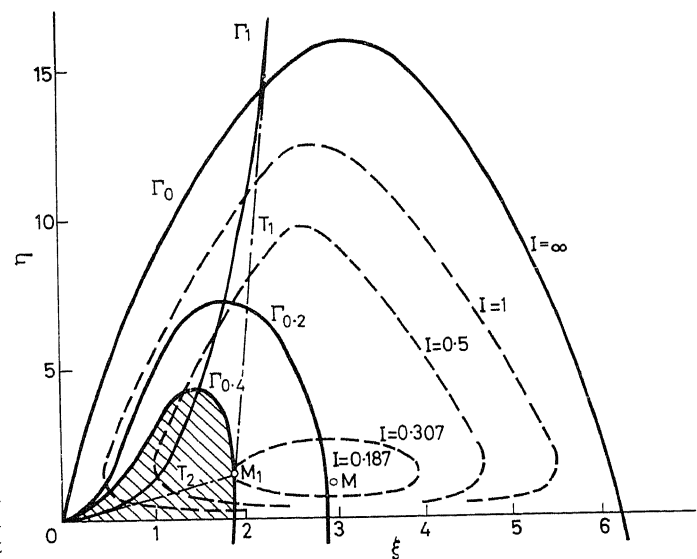


Figure B

technique<sup>1</sup>, one may evaluate the mean squared error  $I$  from eqn (2) as a function of  $\xi$  and  $\eta$

$$I(\xi, \eta) = \frac{0.025\eta^2 - 0.375\xi\eta - 1.55\eta + 10^{-4}\xi - 0.65 \cdot 10^{-3}}{0.625\eta^2 + \xi^2\eta - 6.25\xi\eta} \quad (3)$$

For different values of  $I$ , a family of the curves determined by the eqn (3) is plotted in Figure B.

Now, in order to minimize the mean squared error and simultaneously maintain the prescribed degree of relative stability, the solution of the control problem under investigation is found as a constrained minimum of  $I$ , which is located at the point  $M_1$  of the diagram of Figure B. Within the relative damping region of  $\xi = 0.4$  the point  $M_1$  has minimum value of the mean squared error.

The curve  $I_{0.4}$  and the tangents  $T_1$  and  $T_2$  drawn from the point  $M_1$ , enable the roots of the characteristic equation corresponding to that point to be evaluated without any calculation<sup>2-4</sup>

$$s_{1,2} = -4.1 \pm j9.39; \quad s_3 = -0.864; \quad s_4 = -16.4$$

The smaller real root  $s_3$  and the zero of the corresponding closed-loop transfer function form a dipole whose effect may be neglected. The other real root is large and its effect is also negligible. The step-function response will therefore be governed only by the pair of complex roots  $s_{1,2}$  whose value of damping coefficient  $\xi = 0.4$  ensures that the overshoot is less than 30 per cent.

From the coordinates of the point  $M_1$  ( $\xi = K + 0.01 = 1.89$ ;  $\eta = 0.04$ ;  $K\lambda = 1.52$ ) the values of the compensator parameters are:  $K = 1.88$  and  $\lambda = 20.2$ . The system velocity constant is 38 times greater than the velocity constant of the uncompensated system. The constrained minimum of the mean squared error is  $I = 0.307$ . It is interesting to note from Figure B that the absolute minimum  $I = 0.187$ , which is located at the point  $M$ , falls outside the relative damping region corresponding to  $\xi = 0.2$  and therefore results in a poorly damped system.

In a similar manner<sup>3</sup>, the component of the mean squared error corresponding to the noise can also be expressed as a function of  $\xi$  and  $\eta$ , which for  $\gamma_N = 2\pi$  has the following form

$$J(\xi, \eta) = \frac{\eta\xi - 6.25\eta - 25\xi^2}{0.625\eta + \xi^2 - 6.25\xi} \quad (4)$$

Using this equation, the same reasoning outlined above may be applied to the noise case. Furthermore, by the proposed techniques, it is also possible to introduce the quadratic constraint of the saturation signal into the optimization in the  $\phi\xi\eta$  plane<sup>3</sup>.

## References

- <sup>1</sup> NEWTON, G. C., JR., GOULD, L. A. and KAISER, J. F. *Analytical Design of Linear Feedback Controls*. 1957. New York; Wiley

- <sup>2</sup> ŠILJAK, D. D. Generalization of Mitrović's method, *I.E.E.E. Summer General Meeting, Paper No. 63-988*, 1963, Toronto

- <sup>3</sup> ŠILJAK, D. D. Optimization of squared error with relative stability constraint. *Ph. D. Thesis*, University of Belgrade, Belgrade 1963

- <sup>4</sup> MITROVIĆ, D. Graphical analysis and synthesis of feedback control systems. II-synthesis. *Trans. Amer. Inst. elect. Engrs.* 77(1958)487

R. S. GAYLORD, in reply

I thank Professor Šiljak for his very interesting discussion of my paper. It is particularly interesting to see how he has applied this theory to closed-loop control system design. Although the work that he describes is not directly concerned with dual-input systems, nor is the problem of added constraint on  $x(t)$  treated, it is clear that his topological methods can be applied to these considerations by mapping the constraints into his  $\xi, \eta$  plane.

An additional important observation is that Professor Šiljak must choose a form for his compensation network [see eqn (1) of the discussion], whereas with the approach given<sup>1</sup> this is not necessary. However, if the method fails to find a solution due to this constraint, then another network may be chosen, and so forth. The network found by such a procedure may not be 'optimum' in the sense of the paper, and further, the introduction of the distortionless constraint into Professor Šiljak's method would be very difficult, that is, to satisfy the equation

$$W_1 G_f + W_2 G_f = 1$$

may narrow the region in the  $\xi, \eta$  plane to the point where the solution may not exist.

As implied by Professor Šiljak's opening remarks,  $\sigma_s^2$  is not a very good way to constrain the signal amplitude from saturating. It seems that a more suitable application is to conserve fuel or energy, then the practical choice is much easier. Of course,  $\sigma_s^2$  need not be chosen in the problem, as the trade between  $x^2$  and  $e^2$  may be plotted and compared visually.

## Reference

- <sup>1</sup> NEWTON, G. C., GOULD, L. A. and KAISER, J. F. *Analytical Design of Linear Feedback Controls*. 1957. New York; Wiley

F. CSAKI, Department of Automation, Polytechnic University, Egri 7v 18, v, Budapest XI, Hungary

In his very interesting paper the author has considered a dual-input system with one saturation constraint. However, this problem is only a special case of the multi-variable systems with many constraints. The problem can be generalized as seen in Figure A.

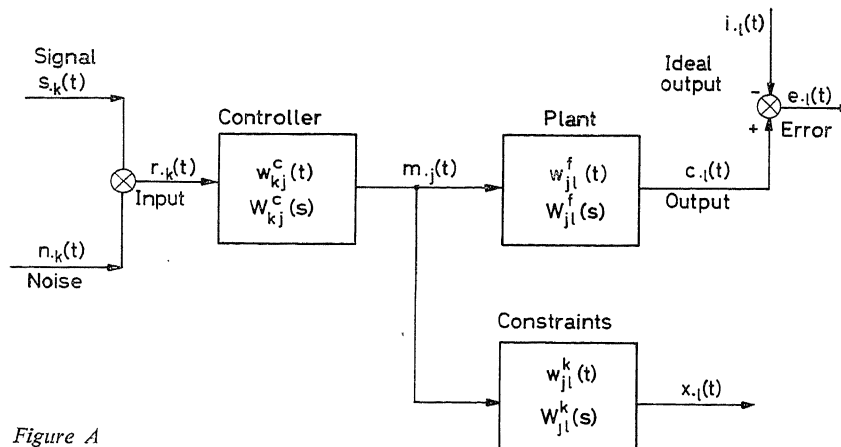


Figure A

Here all lower case letters denote time functions, for example  $s, k, n, k, r, k, m, j, c, l, e, l$ , and  $x, l$  are row vectors, while  $w_{kj}^c, w_{jl}^f$ , and  $w_{jl}^k$  are the weighting-function matrices. The task is to minimize the following function:

$$f^2(t) = \text{tr} \{ \overline{e_l, e_l} + \lambda \overline{x_l, x_l} \} = \text{tr} \{ \phi_{e_l' e_l}(0) + \lambda \phi_{x_l' x_l}(0) \}$$

for stochastic processes by an optimum matrix of the cascade controller  $w_{kj}^{cm}$ , where  $\text{tr}$  denotes the trace of a matrix, i.e. the sum of its diagonal elements.

Instead of correlation functions we can use the power spectral densities:

$$f^2(t) = \frac{1}{2\pi j} \int_{-j\infty}^{j\infty} \text{tr} \{ \phi_{e_l' e_l}(s) + \lambda \phi_{x_l' x_l}(s) \} ds$$

where  $s = j\omega$ . The condition of constraints can be expressed as

$$\frac{1}{2\pi j} \int_{-j\infty}^{j\infty} \text{tr} \{ \phi_{x_l' x_l}(s) \} ds \leq \sigma^2$$

After some mathematical manipulations the following expression can be obtained for the matrix integrand

$$\begin{aligned} & \phi_{e_l' e_l}(s) + \lambda \phi_{x_l' x_l}(s) \\ &= \phi_{i_l' i_l}(s) - \phi_{i_l' r_k}(s) W_{kj}^c(s) W_{jl}^f(s) \\ & - W_{l' j'}^f(-s) W_{j' k'}^c(-s) \phi_{r_k' i_l'}(s) \\ & + W_{l' j'}^f(-s) W_{j' k'}^c(-s) \phi_{r_k' i_l'}(s) W_{kj}^c(s) W_{jl}^f(s) \\ & + \lambda W_{l' j'}^k(-s) W_{j' k'}^c(-s) \phi_{r_k' r_k}(s) W_{kj}^c(s) W_{jl}^c(s) \end{aligned}$$

Let us now introduce an auxiliary power spectral density matrix  $\phi_{a_k, a_k}(s)$  by equating the last two terms on the right-hand side with the following expression.

$$W_{l' j'}^f(-s) W_{j' k'}^c(-s) \phi_{a_k' a_k}(s, \lambda) W_{kj}^c(s) W_{jl}^f(s)$$

Without going into details, the optimizing transfer function matrix of the controller can be expressed in the following explicit matrix relation

$$\begin{aligned} W_{kj}^{cm}(s, \lambda) &= [\phi_{a_k' a_k}^+(s, \lambda)]^{-1} \{ [\phi_{a_k' a_k}(s, \lambda)]^{-1} \phi_{r_k' i_l}(s) W_{l' j'}^f(-s) \\ & \times [(W_{jl}^f(s) W_{l' j'}^f(-s))^{-1}] + [(W_{jl}^f(s) W_{l' j'}^f(-s))^+]^{-1} \} \end{aligned}$$

The solution for  $W_{jk}^{cm}(s, \lambda)$  may now be substituted into the expression

$$\phi_{x_l' x_l}(s, \lambda) = W_{l' j'}^k(-s) W_{j' k'}^c(-s) \phi_{r_k' r_k}(s) W_{kj}^c(s) W_{jl}^k(s)$$

and the Lagrangean variable  $\lambda$  can be adjusted so that the condition of constraint will be satisfied.

These results are the generalizations of some results obtained in the paper under discussion, the details will be published soon in *Periodica Polytechnica, Budapest*. My results obtained by the so-called simplified method are generalizations of the results obtained by Wiener, Amara, Leondes, Newton and others.

R. S. GAYLORD, in reply

I appreciate the important addition made by Professor Csaki in generalizing the problem of multivariable systems with constraints. Due to this valuable addition to the theory one can now consider many signals, controllers, plants and constraints. It is true, however, that the amount of algebraic manipulation greatly increases with the amount of consideration desired. Also, the plant must still be of a very simple form and the signal and noise quantities known in advance, at least as a power spectral density.

The so-called distortionless method eliminates the need to have a known random process as a signal quantity; but it does not appear to be so simple to include this kind of constraint in the more general formulation given.

# Signalling and Prediction of Failures in Discrete Control Devices with Structural Redundancy

M. A. GAVRILOV

## Summary

This report examines the problem of equivalent conversions of the unit structures of discrete control devices whose functioning algorithms are described in the language of conversion tables. An algorithm is described which may be used to convert from any complex structure of a system of units, including those with feedback which require for each conversion an expenditure of  $n\tau$  units of time ( $n$  is the number of units acting in sequence upon each other), to one equivalent unit which requires  $\tau$  units of time to perform the same functioning algorithm. An evaluation is made of the change in structural complexity of such equivalent units. Methods of signalling and predicting failures in discrete control devices with structural redundancy are considered.

## Sommaire

Ce rapport traite des conversions équivalentes de structures unitaires d'éléments de commande discontinue, dont les algorithmes opérationnels sont décrits par des tables de conversion. On montre un algorithme pouvant être utilisé pour convertir, n'importe quelle structure complexe représentant un ensemble d'unités, y compris des unités à asservissement qui demandent pour chaque conversion  $n\tau$  unités de temps ( $n$  est le nombre d'unités agissant en séquence les unes sur les autres), à une structure unitaire équivalente qui demande  $\tau$  unités de temps pour effectuer le même algorithme opérationnel. On évalue le changement de complexité structurale de telles unités équivalentes. On examine des méthodes pour signaler et prévenir les pannes de fonctionnement des éléments de commande discontinue avec redondance structurale.

## Zusammenfassung

Dieser Beitrag untersucht das Problem der gleichwertigen Umwandlung der Elementarstrukturen unstetiger Regelgeräte, deren Funktionsschema mit Hilfe von Umwandlungstabellen beschrieben wird. Es wird ein Verfahren beschrieben, mit dem man jede komplexe Elementarstruktur, darunter auch solche mit Rückführung, die für jede Umwandlung  $n\tau$  Zeiteinheiten brauchen, in ein gleichwertiges Schema umwandeln kann, das  $\tau$  Zeiteinheiten benötigt, um die gleiche Funktionsvorschrift zu erfüllen (dabei ist  $n$  die Anzahl der Einheiten, die im Wirkungsablauf aufeinander folgen). Die Änderungen in der strukturellen Komplexität derartiger gleichwertiger Elemente werden geschätzt. Es werden Verfahren für Anzeige und Vorhersage von Ausfällen in diskreten Regelgeräten mit redundanter Struktur betrachtet.

In solving problems of providing reliable operation of automatic control devices, a great deal of attention is devoted to the use of methods involving the application of structural redundancy. These include all possible methods of duplicating individual elements within units, as well as the more common methods of providing redundancy of all the necessary elements and units on the whole with the least possible number of additional elements. The ever-increasing practical use of methods of structural redundancy is a result of the fact that, in present complex

automatic systems, the control devices require such a large number of individual elements to perform their functions that even though the elements may have a very high reliability, the necessary reliability demanded of the entire device cannot be achieved.

A number of works<sup>1-6</sup> is devoted to the question of the introduction of structural redundancy and the determination of the minimum number of additional elements necessary to achieve the prescribed reliability of the device on the whole. For discrete control devices it is most natural and suitable to examine the required value of operating reliability of the device as being prescribed by a certain number of elements which simultaneously fail during operation while nevertheless permitting the device to perform accurately the control algorithm assigned to it<sup>7</sup>.

The author of the present report showed<sup>3</sup> that when the problem is treated in this manner, the determination of the minimum number of additional internal elements necessary to achieve a given reliability completely coincides with the task of determining the minimum number of additional symbols in the construction of correcting codes with correction of the corresponding number of errors. In the same article a method was given for constructing tables of states which provide for a realization of the structure of a discrete control device having the required reliability.

The proposed method links the problem of constructing such a device to the distribution of the states of its internal elements along the vertices of a many-dimensional cube of single transitions in such a manner that the number of transitions (distance) between the vertices, selected for the corresponding stable states of the device, would be no less than:

$$D = 2d + 1 \quad (1)$$

where  $d$  is the number of simultaneously failing elements with which the devices must still exactly perform their control algorithm.

In differentiating the demands on reliability (namely, separating them from the viewpoint of the number of simultaneously failing elements), first, into that for which the device must accurately perform a given control algorithm and, second, into that for which it must not provide any actions at its outputs, the value of the distance between vertices selected for the stable states must be no less than:

$$D = 2d + \Delta + 1 \quad (2)$$

where  $\Delta$  is the number of simultaneously failing elements in addition to  $d$  for which the indicated second condition of reliable operation of the device must be fulfilled.

In discrete types of devices which have reliability as a result of structural redundancy, the required reliability is retained

only until the moment of onset of permanent failure of even one of the elements.

In fact, let the prescribed probability of failure of the entire device on the whole require that the given control algorithm be exactly performed with the simultaneous failure of  $d$  elements. Then, with a permanent failure of any one of the elements, the device will capably perform the control algorithm only upon the simultaneous failure of  $d - 1$  elements; that is, it will have a probability of complete failure which is less than prescribed.

Particular importance is therefore devoted to rapid signalling of failure of individual elements or their prediction, which permits one to take timely measures to replace the faulty elements or other measures which will return the probability of failure of the entire device to its prescribed value. The present report is devoted to an examination of the fundamental possibilities of providing such signalling or prediction for automatic control devices designed on the basis of the principles described by the author<sup>3</sup>.

First it is shown that the table of states constructed according to the principles contains all the necessary information on failure, both generally for all elements as well as for each of them individually, and, even more, on the nature of the failures.

Those states of internal elements which correspond to the stable states of the corresponding table of transitions and which are distributed, as was pointed out above, in the vertices of a many-dimensional cube of single transitions with a distance one from the other of not less than  $D$ , are called basic. To each of these states there must correspond a particular state of outputs which provides for the performance of the prescribed control algorithm.

Let the number of inputs of the discrete device be equal to  $a$  and let it be given that, to perform the control algorithm with a prescribed degree of reliability, that is, in the presence of simultaneous failure of  $d$  internal elements, it is necessary to have  $K$  internal elements. Then each of the basic states may be characterized by a certain conjunctive member of a Boolean function of length  $a + K$ . In accordance with this the table of states contains, on the left-hand side,  $a + K$  columns of which  $a$  characterizes the states of the inputs and  $K$  characterizes the states of the internal elements. The binary number characterizing the state of the internal elements corresponds to a particular vertex of the many-dimensional cube, selected in distributing the given basic state.\*

The failure of any element is characterized by a change in the binary number, corresponding to a given basic state, from zero to one or one to zero. The first is called a  $0 \rightarrow 1$  type failure and the second a  $1 \rightarrow 0$  type failure. Each such failure transfers the basic state to an adjacent vertex of the many-dimensional cube. The simultaneous failure of any two internal elements transfers the basic state to a vertex two units removed from the vertex selected for the given basic state; it is adjacent to any vertex to which the basic state was transferred by the failure of any one of these two elements.

In order to provide exact performance of the control algorithm upon the failure of internal elements, each of the states to which the basic state is transferred upon the failure of any number of elements within the prescribed limits (that is, inclusive to  $d$ ) must compare in the right-hand side of the table

of states to the same state of outputs as the basic state. Therefore, for each stable state of the table of transitions, for the case of structural redundancy, there must correspond a particular combination of states consisting of the basic state and all the states to which it transfers upon failure of the internal elements. All of these states are adjacent to one another, forming a certain multiple of adjacent states. This multiple is called a set of basic states.

First it is shown that the set of adjacent states, together with the basic states, may be described by a symmetrical Boolean function whose active numbers represent a natural series of numbers from  $K - d$  to  $K$ .

Let there be any state  $f_{i0}$  corresponding to one of the basic states and let this state be characterized by a row in the table of states containing  $K_1$  zeros and  $K_2$  ones, where  $K_1 + K_2 = K$ . Then, with  $d = 1$ , the collection of adjacent states  $\Sigma f_{i1}$  contains all the states differing from the basic by the replacement of one variable by its reciprocal. More precisely, they are  $K$ , while  $K_1$  of them corresponds to a failure of the type  $0 \rightarrow 1$  and  $K_2$  to a failure of the type  $1 \rightarrow 0$ . It is easy to see that the sum of these states may be characterized by the symmetrical function:

$$\Sigma f_{i1} = S_{K-1}(\bar{x}_1, \bar{x}_2, \dots, \bar{x}_{K_1}, \bar{x}_{K_1+1}, x_{K_1+2}, \dots, x_{K_1+K_2})$$

if the basic state is considered a symmetrical function of those variables with an active number equal to  $K$ , namely:

$$f_{i0} = S_K(\bar{x}_1, \bar{x}_2, \dots, \bar{x}_{K_1}, x_{K_1+1}, x_{K_1+2}, \dots, x_{K_1+K_2})$$

The sum of the basic and set of adjacent states is thus characterized by the symmetrical Boolean function:

$$f_{i0} + \Sigma f_{i1} = S_{K-1, K}(\bar{x}_1, \bar{x}_2, \dots, \bar{x}_K, x_{K+1}, x_{K+2}, \dots, x_{K_1+K_2})$$

If  $d = 2$ , the set of adjacent states consists of all states differing from the basic by the replacement of one variable by its reciprocal, the number of which, as was pointed out, is equal to  $K = C_K^1$ , and two variables. The number of the latter is obviously equal  $C_K^2$ , and since each of them differs from the basic by a change having a value of two variables, their total  $\Sigma f_{i2}$  corresponds to the symmetrical function:

$$\Sigma f_{i2} = S_{K-2}(\bar{x}_1, \bar{x}_2, \dots, \bar{x}_{K_1}, \bar{x}_{K_1+1}, x_{K_1+2}, \dots, x_{K_1+K_2})$$

The Boolean function characterizing the basic state and the entire set of adjacent states is thus a symmetrical function of the type:

$$\begin{aligned} f_{i0} + \Sigma f_{i1} + \Sigma f_{i2} \\ = S_{K-2, K-1, K}(\bar{x}_1, \bar{x}_2, \dots, \bar{x}_{K_1}, x_{K_1+1}, x_{K_1+2}, \dots, x_{K_1+K_2}) \end{aligned}$$

It may be proved in an analogous manner that in the general case, with the simultaneous failure of  $d$  internal elements, the basic state and the set of adjacent states may be characterized by a symmetrical Boolean function of the type:

$$S_{K-d, K-d+1, \dots, K}(\bar{x}_1, \bar{x}_2, \dots, \bar{x}_{K_1}, x_{K_1+1}, x_{K_1+2}, \dots, x_{K_1+K_2})$$

Thus, the class of reliable structures of discrete devices is, with respect to internal elements, a class described by symmetrical Boolean functions of a special type, which facilitates their realization since these functions have been most widely studied and may be economically realized with the aid of different types of threshold relay elements, including electromagnetic relay elements with several windings<sup>8</sup>.

\* All references made below to internal elements with an identical base pertain to inputs and sensing elements.

The basic state is designated as  $f_i$  and the set of adjacent states corresponding to it as  $N_i$ , assuming that  $f_i + N_i = F_i$ .

The table of states of a discrete control device consists on the left-hand side of all sets  $F_i$  combined with the corresponding values of inputs. For each of these sets there corresponds on the right-hand side of the table, as was pointed out above, a state of outputs which provides for the performance of the control algorithm. One more output is added for which is included in the table of states a zero for each of the basic states and a one for any of the states which are included in the sets of adjacent states.

Since the latter corresponds to the failure of any one or to the simultaneous failure of several internal elements, the appearance of a one at this output occurs only by means of a decrease in the reliability of operation of the discrete device and may be used to signal the presence of a failure.

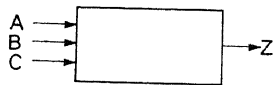


Figure 1

For example, let there be a discrete device with three inputs and one output (Figure 1) and an action, equal to one, must appear at the latter in the subsequent sequence of change of the states of the outputs:

0 0 0  
1 0 0  
1 1 0  
1 1 1  
0 1 1

Any subsequent change of inputs must lead to the appearance of an action at the output equal to zero, while the further appearance of an action at the output equal to one occurs only by the repetition of the indicated sequence of change of the states of the inputs. With any other sequence of change of the states of the inputs, the action at the output must remain equal to zero.

The corresponding table of conversions is given in Table 1. Here it may be seen that it is necessary to provide for four stable states, which is possible with the aid of two internal elements.

When it is necessary that the aforementioned discrete device performs exactly a preassigned control algorithm in the event of the simultaneous failure of one of the internal elements, five

Table 1

000	100	110	010	011	111	101	001
(1) <sup>0</sup>	(1) <sup>0</sup>	2	4	(1) <sup>1</sup>	4	4	4
—	4	(2) <sup>0</sup>	4	—	3	—	—
—	—	4	—	1	(3) <sup>0</sup>	4	—
1	(4) <sup>0</sup>	(4) <sup>0</sup>	(4) <sup>0</sup>	(4) <sup>0</sup>	(4) <sup>0</sup>	(4) <sup>0</sup>	(4) <sup>0</sup>

internal elements are required, as seen in Table 5 of reference 3.

The following distribution for the basic states is chosen:

0 0 0 0 0  
1 0 1 1 0  
0 1 0 1 1  
1 1 1 0 1

Then the table of states will have the form shown in Table 2. In agreement with what was mentioned above, let us add the

output  $C_0$ , in the column of which are written zeros in all the rows of the table of states corresponding to  $f_i$  and ones in all the rows corresponding to  $N_i$  (Table 3). Then this output will signal the presence of a failure of any one or several of the internal elements.

Table 2

A	B	C	F	$X_1$	$X_2$	$X_3$	$X_4$	$X_5$	Z
0	0	0	$F_2$	0	0	0	0	0	0
0	0	0	$F_4$	0	0	0	0	0	0
0	0	1	$F_1$	1	1	1	0	1	0
0	0	1	$F_4$	1	1	1	0	1	0
0	1	0	$F_1$	1	1	1	0	1	0
0	1	0	$F_2$	1	1	1	0	1	0
0	1	0	$F_4$	1	1	1	0	1	0
0	1	1	$F_1$	0	0	0	0	0	1
0	1	1	$F_3$	0	0	0	0	0	1
0	1	1	$F_4$	1	1	1	0	1	0
1	0	0	$F_1$	0	0	0	0	0	0
1	0	0	$F_2$	1	1	1	0	1	0
1	0	0	$F_4$	1	1	1	0	1	0
1	0	1	$F_1$	1	1	1	0	1	0
1	0	1	$F_3$	1	1	1	0	1	0
1	0	1	$F_4$	1	1	1	0	1	0
1	1	0	$F_1$	1	0	1	1	0	0
1	1	0	$F_2$	1	0	1	1	0	0
1	1	0	$F_3$	1	1	1	0	1	0
1	1	0	$F_4$	1	1	1	0	1	0
1	1	1	$F_1$	1	1	1	0	1	0
1	1	1	$F_3$	0	1	0	1	1	0
1	1	1	$F_2$	0	1	0	1	1	0
1	1	1	$F_4$	1	1	1	0	1	0

In this table:

$F_1$	0 0 0 0 0	$F_2$	1 0 1 1 0	$F_3$	0 1 0 1 1	$F_4$	1 1 1 0 1
	1 0 0 0 0		0 0 1 1 0		1 1 0 1 1		0 1 1 0 1
	0 1 0 0 0		1 1 1 1 0		0 0 0 1 1		1 0 1 0 1
	0 0 1 0 0		1 0 0 1 0		0 1 1 1 1		1 1 0 0 1
	0 0 0 1 0		1 0 1 0 0		0 1 0 0 1		1 1 1 1 1
	0 0 0 0 1		1 0 1 1 1		0 1 0 1 0		1 1 1 0 0

If one places the action from this output into a computer and determines the number of times that actions equal to one appear at this output during a certain time interval, the answers from the computer may be used to predict an approximation of reliable operation of the device.

The described principle of signalling and prediction has significant advantages in the sense that neither the signalling nor prediction requires the introduction of any additional internal elements. Usually the performance of these functions relies upon special units of the discrete device which require elements having, in principle, a reliability as much as one order of magnitude greater than the elements which make up the discrete device itself.

In the design examined above, comprising a structure of signal outputs based on actuating devices already having internal elements, and assuming that the connections between these



Table 3

$X_1$	$X_2$	$X_3$	$X_4$	$X_5$	$C_0$	$C_1$	$C_2$	$C_3$	$C_4$	$C_5$
0	0	0	0	0	0	0	0	0	0	0
1	0	0	0	0	1	1	0	0	0	0
0	1	0	0	0	1	0	1	0	0	0
0	0	1	0	0	1	0	0	1	0	0
0	0	0	1	0	1	0	0	0	1	0
0	0	0	0	1	1	0	0	0	0	1
1	0	1	1	0	0	0	0	0	0	0
0	0	1	1	0	1	1	0	0	0	0
1	1	1	1	0	1	0	1	0	0	0
1	0	0	1	0	1	0	0	1	0	0
1	0	1	0	0	1	0	0	0	1	0
1	0	1	1	1	1	0	0	0	0	1
0	1	0	1	1	0	0	0	0	0	0
1	1	0	1	1	1	1	0	0	0	0
0	0	0	1	1	1	0	1	0	0	0
0	1	1	1	1	1	0	0	1	0	0
0	1	0	0	1	1	0	0	0	1	0
0	1	0	1	0	1	0	0	0	0	1
1	1	1	0	1	0	0	0	0	0	0
0	1	1	0	1	1	1	0	0	0	0
1	0	1	0	1	1	0	1	0	0	0
1	1	0	0	1	1	0	0	1	0	0
1	1	1	1	1	1	0	0	0	1	0
1	1	1	0	0	1	0	0	0	0	1

devices and the sensing signal and predicting devices have 100 per cent reliability, one would expect that the signalling of failure would have absolute reliability in principle.

In fact, only two mutually exclusive events may occur: (a) not one of the internal elements is faulty. Then the actions equal to one appear at the corresponding operating outputs and at the signal output the action is equal to zero; (b) failure of one or several internal elements occurs within the limits of  $d$ . Then an action equal to one appears both at the signal and operating outputs.

It is noted that achieving reliable operation by means of the introduction of structural redundancy according to the principles previously presented by the author<sup>3</sup> pertain to the internal elements of the device as a whole, that is, both to the actuating and the reacting devices. Therefore, with respect to failures of the actuating organs, the device retains its ability to perform exactly the control algorithm upon the failure of either one or, simultaneously, all of the actuating devices of a given internal element for the conditions when these failures are all of a single type.

The described principle of designing signal circuits makes it possible to provide separately for signalling the number of failures greater than  $d$ , including those located between the limits of  $d + 1$  and  $d + \Delta$ . Additional outputs must be added for this purpose. This requires that ones be written in the specific rows in the appropriate columns of the table of states; namely, for signalling failures of elements within limits from  $d + 1$  to  $d + \Delta$  in the rows corresponding to failures in these limits, and for signalling a large number of failures in the rows corresponding to unused states.

It is obvious that the signalling of failures may be not only general but also specific, or, for each of the internal elements of the device separately. For this purpose one must have for each of them an individual output, for which there must be written in the columns of the table of states *ones* for all states differing from the basic by the change in value of the corresponding variable. For example, to signal the failure of element  $X_1$  in the above case, ones must be written for each first row of the sets  $N_i$  for the corresponding output.

Table 3 gives the corresponding values of outputs for each of the internal elements. The realization of such outputs provides, in the event of faulty elements in the device, for advance notification as to which of the internal elements is malfunctioning or, with prediction, an approximate indication, permitting timely replacement or adjustment of the element for proper action.

Obviously it is possible to provide not only for signalling of failures of individual internal elements but for the separate signalling of the nature of these failures as well. For example, in Table 4, for the internal element  $X_1$  examined above, are shown the operating states corresponding to failures of the type  $0 \rightarrow 1$  [Table 4(a)] and failures of the type  $1 \rightarrow 0$  [Table 4(b)].

Table 4

$X_1$	$X_2$	$X_3$	$X_4$	$X_5$	$X_1$	$X_2$	$X_3$	$X_4$	$X_5$
1	0	0	0	0	0	0	1	1	0
1	1	0	1	1	0	1	1	0	1

(a)

(b)

In conclusion some of the problems of realizing signalling and prediction networks are considered. The circuit of each output in the structure of a multi-cycle discrete device must contain actuating devices of both internal and sensing relay elements. The signal circuits must contain actuating devices of only internal elements. Therefore the rational design of the structure of a discrete device would be that shown in Figure 2, namely, a structure in the form of a certain  $[1, K]$  terminal network having at its outputs all the functions of  $f_i$  and  $N_i$  and containing the actuating devices of only the internal elements, and an  $[M, N]$  terminal network containing the actuating devices of only the sensing elements.

As was pointed out above, the functions which realize the basic states together with the sets of adjacent states are symmetrical with the operating numbers from  $K - d$  to  $K$  and for their realization it is suitable to use so-called 'threshold' elements. When such elements are used it is advantageous to use the structure of the discrete device having a form shown in Figure 2(b), where the  $[1, K]$  terminal network is based on threshold elements according to the number of basic states. The  $[M, N]$  terminal network has the same make-up as that shown in Figure 2(a), while the output circuits for signalling and prediction of failures are derived from the outputs of the threshold elements by means of their series connection (providing an 'and' operation) and from circuits corresponding to the function  $\bar{f}_i$ . The latter may also be designed with the aid of threshold elements having symmetrical functions with the operating number  $K$ .

In addition it is noted that, in the case examined above, it is most rational from the viewpoint of the simplest physical realization of the structure of a discrete device to choose the operating levels of the symmetrical functions not from  $K - d$

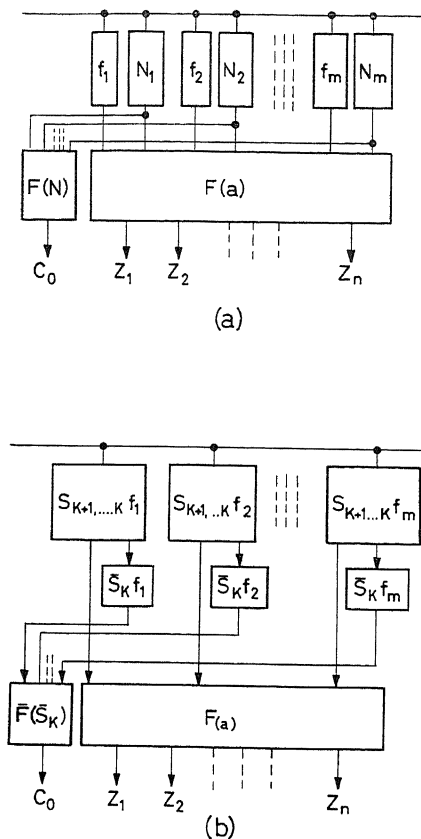


Figure 2

to  $K$  but from 0 to  $d$ , while simultaneously taking not the variables but their inversions.

In conclusion one should note that the method considered previously by the author<sup>3</sup>, as well as everything discussed in this report, refer to the case in which the probability of failure for all internal elements has a single value, the failures are symmetrical (that is, the probability of failures of the type  $0 \rightarrow 1$  is identical to that of type  $1 \rightarrow 0$ ), and, in addition, failures of individual elements are mutually independent. Conditions differing from these necessitate a somewhat different approach to determining the minimum number of elements and the distribution of the states. However, the principles of designing signal circuit and of prediction remain the same, with the exception that the functions characterizing the basic sets and the sets of adjacent states may not prove symmetrical.

### Further Work

The problem considered in the paper deals particularly with the signalling of failures in the so-called delay unit of discrete controllers ( $X$  in Figure 3). It is a little more complicated to signal failures in the combinational part of the device ( $Q$  in

Figure 3). In the papers published up to now it was proposed as a solution to use special signalling circuits composed of elements whose reliability is as much as one order higher than the reliability of the main elements. These circuits have the complexity and sometimes even more, in comparison with the main circuit.

The method considered in the paper permits the use, in the signalling circuits, of the same elements as were used in the main circuits in case of absolute coupling between actuating and reacting parts of the relay elements. On the other hand, the absence of such a coupling allows considerable reduction in the

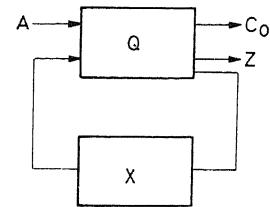


Figure 3.  $A$  = Inputs;  $Z$  = Outputs;  $C_0$  = Outputs for transmitting signals;  $Q$  = Combinational parts of device;  $X$  = Discrete controllers

complexity of signalling circuits and improves reliability. It would be useful, in this case, to apply the method of combinational circuit design proposed by ZAKREVSKY<sup>9</sup>, whose method provides signalling of failures by means of a combination of actions at the output of discrete control systems.

### References

- <sup>1</sup> VON NEUMAN, S. Probabilistic logics and the synthesis of reliable organisms from unreliable components. *Automata Studies*. 1956. Princeton; Princeton University Press
- <sup>2</sup> MOORE, E. F. and SHANNON, C. E. Reliable circuits using less reliable relays. *J. Franklin Inst.* Vol. 262, No. 3 (1956) 191, 281
- <sup>3</sup> GAVRILOV, M. A. Structural redundancy and reliability of relay circuits. *Automatic and Remote Control*. Vol. 2, p. 838. 1961. London; Butterworths
- <sup>4</sup> ZAKROVSKIY, A. D. A method of synthesis of functionally stable automata. *Dok. AN SSSR* Vol. 129, No. 4 (1959) 729
- <sup>5</sup> RAY-CHANDHURI, D. K. On the construction of minimally redundant reliable system designs. *B.S.T.J.* Vol. 40, No. 2 (1961) 595
- <sup>6</sup> ARMSTRONG, D. B. A general method of applying error correction to synchronous digital systems. *B.S.T.J.* Vol. 40, No. 2 (1961) 577
- <sup>7</sup> GAVRILOV, M. A. Basic terminology of automatic control. *Automatic and Remote Control*. Vol. 2, p. 1052. 1961. London; Butterworths
- <sup>8</sup> GAVRILOV, M. A. *The Structural Theory of Relay Devices, Part 3. Contactless Relay Devices*. 1961. Moscow; Publishing House of the All Union Correspondence Power Engineering Institute
- <sup>9</sup> ZAKREVSKY, A. Theory of computations, automation, information theory. *Proc. Siberian Ins. Phys.* 1961. University of Tomsk. p. 12

## DISCUSSION

G. C. MOISIL, *Academy of the R.P.R., Calea Victoriei 125, Bucarest, Rumania*

We are now working on the problem of finding a generalization of Professor Gavrilov's results for the case of many-valued Lukasievicz algebras and for the case of the ring of integers modulo  $n$ .

In fact, polarized relays, latching relays and the real operation of ordinary relays, when the intermediate states are considered, reflect to three-value Lukasievicz algebras.

Certain special relays introduce the  $n$ -valued Lukasievicz algebra. The rotary switch introduces the ring of integers modulo  $n$ . We hope to be able to generalize Professor Gavrilov's results in all these cases.

CHUAN-SHAN WANG, *Institute of Automation, Academy of Sciences, Peking, China*

Having read Professor Gavrilov's paper, which concerns the new and important problem of the construction of relay circuits with given reliability, I make the following remarks.

Due to the hazard of relay operation the circuit proposed by Professor Gavrilov may have intermediate states which can greatly reduce the reliability. For example, the four basic states chosen in the example in the paper are

0	0	0	0	0
1	0	1	1	0
0	1	0	1	1
1	1	1	0	1

When a transition from the second to the third of these states takes place, the intermediate state 1 1 1 1 1 may occur and this state belongs to the reception region of another basic state, i.e., 1 1 1 0 1 (the distance between 1 1 1 1 1 and 1 1 1 0 1 being only 1). If one relay is damaged beforehand, then the situation will be still worse. This phenomenon was first noticed by one of the workers in our laboratory. In order to avoid these effects special methods such as the introduction of time delays or the use of clock pulses, have to be applied, otherwise the arrangement of the basic states, i.e. the construction of the error correcting code, has to be considered very carefully. This will introduce theoretical and practical difficulties, especially when the number of relays is large, and when, at the same time, effectiveness and minimization are considered. What is the author's opinion about these problems?

The second question concerns the practical application of prediction. The author's ideas of prediction of the failure of relays is very interesting. If the arrangement of states is uniform and the reception region of each basic state is symmetrical, a decision element may be used as decoder and the circuit will be very simple. If, however, the circuit is not symmetrical the decoder may be very complex. Since a computer must be used in any case, the practical applicability of the prediction method is not very clear.

Finally, I would mention that in one of my papers which was published in the Chinese journal *Automation* in 1958, I reported on a study of the symmetry characteristics of error-correcting codes and

the construction of decoders. The main points of this paper are very similar to those in the first part of Professor Gavrilov's paper.

M. A. GAVRILOV, *in reply*

Answering Mr. Wang's first point on the question of instability of inner elements during transitions from one stable state to another (hazard phenomena), the suggestion of using relay elements with different time delays is not the best one. The problem is to secure elementary transfers between sets of states (Figure A) where  $f_{ij}$  ( $j = 1, 2, \dots, n$ ) is a basic state and  $N_{ij}$  ( $j = 1, 2, \dots, n$ ) are adjacent states. There are no critical hazards within these sets. They must be considered only between the sets.

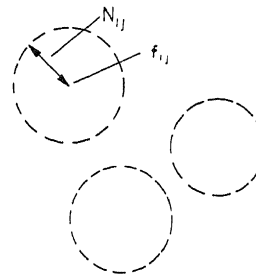


Figure A

The problem might be solved by the known method of assignment of states based on the table of typical assignments suggested by Professor Hoffmann (U.S.A.) which provides for independence of instability of element delays.

The second point deals with the complication of realization of the structure of a device in the case where symmetrical functions cannot be realized with a single inner element. Principally there are no difficulties in the structural synthesis of the combinational part of the relay device, which has a certain reliability; see Gavrilov<sup>1</sup> and Zakrevsky<sup>2</sup>.

In fact the main problem of signalling failure is the reliability of the signalling device, which should be an order of magnitude better than that of the combinational part of the system. The method suggested in the report is decreasing the complexity of the signal circuits and thus making the solution easier.

Unfortunately I am not familiar with Mr. Wang's article which was published in *Automation* in 1958, and can only state that I am satisfied that the results gained proved his conclusions.

## References

- <sup>1</sup> GAVRILOV, M. A. Structural redundancy and reliability of relay circuits. *Automatic and Remote Control*. Vol. 2, p. 838. 1961. London; Butterworths
- <sup>2</sup> ZAKREVSky, H. Functional stability of relay circuits. *Trud. sib. Fiz.-tekhn. Inst. Vychis. Litelnoj Teknika, Automatica, Teory Informacii* (1961)

# Axiomatization of the Theory of Simplification of Combinational Automata

G. R. C. MOISIL

## Summary

The paper shows that the classical method, due to Quine and others, of simplification of dipoles with contacts, is a consequence of axioms I—V and may therefore be applied not only to dipoles with series or parallel contacts and to circuits with diodes, but also to circuits with triodes, transistors or cryotrons.

An example is given in section VII. Upon replacing axiom IV by IV\* the method may be applied to circuits with multivalued elements, e. g. relays with real time functioning, polarized relays, relays with multipositional armatures and rotary switches.

## Sommaire

On montre que la méthode classique de simplification des dipôles avec contact, de Quine et autres, découle des axiomes I—V. En conséquence, elle peut être utilisée non seulement dans le cas des dipôles avec contacts en série ou en parallèle et dans le cas des circuits avec diodes, mais aussi dans le cas des circuits avec triodes, transistors ou cryotrons.

On donne un exemple dans la section VII. En remplaçant l'axiome IV par IV\*, la méthode peut être appliquée aux circuits avec éléments à valeurs multiples avec relais fonctionnant en temps réel, relais polarisés, relais avec armature à positions multiples et commutateurs rotatifs.

## Zusammenfassung

Dieser Beitrag zeigt, daß die klassische Methode der Vereinfachung von Dipolen (nach Quine und anderen) mit Kontakten auf den 5 hier angeführten Axiomen beruht. Sie kann daher nicht nur auf Dipole mit Serien- oder parallelen Kontakten und auf Schaltkreise mit Dioden angewendet werden, sondern auch auf Kreise mit Trioden, Transistoren oder Cryotronen.

Im Abschnitt VII wird ein Beispiel angegeben. Ersetzt man das Axiom IV durch das Axiom IV\*, so kann diese Methode auf Kreise mit mehrwertigen Elementen angewendet werden. Hierzu zählen z. B. Relais mit realem Zeitverhalten polarisierte Relais, Relais mit Mehrstellungsanker und Wahlschalter.

It is intended to derive a calculus which the axioms have to satisfy in order that the simplifying method, given by Quine for the II-dipoles with contacts and relays, should be valid. Quine's method has been thoroughly investigated by many researchers whose contributions are important, especially J. McCluskey and J. Paul Roth. The axioms given here show that this method may be used in many other instances, such as that of circuits with triodes, transistors, cryotrons, or with three positional relays (as in the case of polarized relays or of real operation of ordinary relays) and with multipositional contacts (as in the case of selectors and codified relays).

A. The researches of Quine and of his successors are related to the two expansions in Boole series

$$f(x_1, \dots, x_n) = \bigcup_{\alpha} f(\alpha_1, \dots, \alpha_n) L_{\alpha_1}(x_1) \dots L_{\alpha_n}(x_n) \quad (1)$$

$$f(x_1, \dots, x_n) = \prod_{\alpha} [f(\bar{\alpha}_1, \dots, \bar{\alpha}_n) \cup L_{\alpha_1}(x_1) \cup \dots \cup L_{\alpha_n}(x_n)] \quad (2)$$

where

$$L_0(z) = \bar{z}, \quad L_1(z) = z \quad (3)$$

The formulae are of the following type:

$$f(x_1, \dots, x_n) = \Omega_{\alpha} [c_{\alpha_1, \dots, \alpha_n} \theta L_{\alpha_1}(x_1) \theta \dots \theta L_{\alpha_n}(x_n)] \quad (4)$$

where  $\theta$  and  $\omega$  represent two operations with any number whatever of variables and where expressions such as

$$\begin{aligned} \Omega_{i=1}^r z_i &= z_1 \omega \dots \omega z_r \\ \Theta_{i=1}^r z_i &= z_1 \theta \dots \theta z_r \end{aligned} \quad (5)$$

with  $r > 1$  have a meaning, namely eqns (1) and (2) are eqn (4) if

	$\omega$	$\theta$	$c_{\alpha_1, \dots, \alpha_n}$	
I	$\cup$	$\cdot$	$f(\alpha_1, \dots, \alpha_n)$	(6)
II	$\cdot$	$\cup$	$f(\bar{\alpha}_1, \dots, \bar{\alpha}_n)$	

It has already been shown that six other expansion formulae of the eqn (4) type are valid:

	$\omega$	$\theta$	$c_{\alpha_1, \dots, \alpha_n}$	
III	$\cup$	$\top$	$f(\bar{\alpha}_1, \dots, \bar{\alpha}_n)$	
IV	$\Pi$	$\perp$	$f(\alpha_1, \dots, \alpha_n)$	
V	$\perp$	$\cup$	$f(\bar{\alpha}_1, \dots, \bar{\alpha}_n)$	
VI	$\top$	$\Pi$	$f(\alpha_1, \dots, \alpha_n)$	(7)
VII	$\perp$	$\perp$	$f(\alpha_1, \dots, \alpha_n)$	
VIII	$\top$	$\top$	$f(\bar{\alpha}_1, \dots, \bar{\alpha}_n)$	

The functions  $\top$  and  $\perp$  are Sheffer's functions of several variables

$$\begin{aligned} z_1 \top \dots \top z_r &= \bar{z}_1 \dots \bar{z}_r \\ z_1 \perp \dots \perp z_r &= \bar{z}_1 \cup \dots \cup \bar{z}_r \end{aligned} \quad (8)$$

The interpolation formula of Lagrange in GF(2) is of the same type; as a matter of fact there are here two Lagrange interpolation formulae for GF(2) (IX and its dual X):

	$\omega$	$\theta$	$c_{\alpha_1, \dots, \alpha_n}$	
IX	$+$	$\cdot$	$f(\alpha_1, \dots, \alpha_n)$	(9)
X	$\nabla$	$\cup$	$f(\bar{\alpha}_1, \dots, \bar{\alpha}_n)$	

The functions  $+$  and  $\sim$  of several variables are defined recurrently

$$\begin{aligned} z_1 + \dots + z_n &= (z_1 + \dots + z_{n-1}) + z_n \\ z_1 \sim \dots \sim z_n &= (z_1 \sim \dots \sim z_{n-1}) \sim z_n \end{aligned} \quad (10)$$

while

$$\begin{aligned} \alpha + \beta &= \alpha \bar{\beta} \cup \bar{\alpha} \beta \\ \alpha \sim \beta &= \alpha \beta \cup \bar{\alpha} \bar{\beta} \end{aligned} \quad (11)$$

In all these cases,  $f(\alpha_1, \dots, \alpha_n)$  is 0 or 1 and eqn (4) reduces itself to the function generated by an expression

$$f(x_1, \dots, x_n) = f_{\mathcal{C}}(x_1, \dots, x_n) \quad (12)$$

where  $\mathcal{C}$  is an expression, yielding the definition:

I. The expressions are sequences of letters of the following form

$$\bigcap_{h=1}^r (z_{h1} \theta \dots \theta z_{hm_h}) \quad (13)$$

If  $r > 1$ ,  $\bigcap_{h=1}^r$  is defined by eqn (4). If  $r = 1$ ,  $\bigcap_{h=1}^r t_h$  is  $t$  if  $\omega$  is  $\cup, \cdot, +, \sim$  and  $\bar{t}$  if  $\omega$  is  $\cap, \vee, \perp$ ; in (13),  $z_{ij}$  will be substituted by  $L_{\alpha}(x_{\beta})$  and therefore by  $x_{\beta}$  or  $\bar{x}_{\beta}$ .

II. Between the expressions, a relation of equivalence  $=$  may take place, satisfying the following conditions if

$$\mathcal{A}_i = \mathcal{D}_i$$

then

$$\begin{aligned} \bigcap_i \mathcal{A}_i &= \bigcap_i \mathcal{D}_i \\ \bigcap_i \mathcal{A}_i &= \bigcap_i \mathcal{D}_i \end{aligned}$$

Evidently,  $\cup, \dots, \sim$  satisfy condition II.

B. The application of Quine's method is based on the formulae

$$\begin{aligned} x \cup \bar{x} &= 1 \\ 1 \cdot x &= x \end{aligned} \quad (14)$$

Therefore

$$\begin{aligned} z_0 z_1 \dots z_r \cup \bar{z}_0 \bar{z}_1 \dots \bar{z}_r \cup t_1 \cup \dots \cup t_s \\ = (z_0 \cup \bar{z}_0) z_1 \dots z_r \cup t_1 \cup \dots \cup t_s \\ = 1 z_1 \dots z_r \cup t_1 \cup \dots \cup t_s \\ = z_1 \dots z_r \cup t_s \cup \dots \cup t_s \end{aligned} \quad (15)$$

In order to apply this method, the terms  $z_0 z_1 \dots z_r, \bar{z}_0 \bar{z}_1 \dots \bar{z}_r$  must be brought to be neighbours and the variables must be arranged in a definite order. Therefore, it will be assumed that.

III.  $\Omega$  and  $\Theta$  are commutative; that is to say if  $\pi$  is a permutation of the indexes  $1, \dots, r$ , then

$$\begin{aligned} z_{\pi(1)} \theta \dots \theta z_{\pi(r)} &= z_1 \theta \dots \theta z_r \\ z_{\pi(1)} \omega \dots \omega z_{\pi(r)} &= z_1 \omega \dots \omega z_r \end{aligned} \quad (16)$$

This property allows the expression to take the form

$$(z_0 \theta z_1 \theta \dots \theta z_r) \omega (\bar{z}_0 \theta z_1 \theta \dots \theta z_r) \omega t_1 \omega \dots \omega t_s$$

The commutativity is valid for all the operations given as examples:  $\cup, \cdot, \cap, \vee, +, \sim$ . Yet, in order to make the simplification, it is not necessary that all the steps in eqn (15) could be made. It is sufficient that

IV. The following equality be true

$$\begin{aligned} (z_0 \theta z_1 \theta \dots \theta z_r) \omega (\bar{z}_0 \theta z_1 \theta \dots \theta z_r) \omega t_1 \omega \dots \omega t_s \\ = (z_1 \theta \dots \theta z_r) \omega t_1 \omega \dots \omega t_s \end{aligned} \quad (17)$$

It is important to emphasize that for all the pairs of operations  $\omega, \theta$ , from eqns (6), (7) and (9), eqn (17) remains true. That is so much more remarkable, since the various steps made in eqn (15), such as the associativity of  $\cup$ , the distributivity of  $\cdot$ , with regard to  $\cup$ , etc. are not valid for some of these pairs (I-X) of  $\omega, \theta$  operations.

C. This first stage of simplification is valid for:

(a) the dipoles  $\Pi$  with contacts, as well in the normally disjunctive form ( $\omega = \cup, \theta = \cdot$ ) as in the normally conjunctive form ( $\omega = \cdot, \theta = \cup$ );

(b) the diode circuits, in the same cases;

(c) the triode circuits, of the two following forms

$$\omega = \cup, \quad \theta = \cap \quad (\text{form III})$$

$$\omega = \cap, \quad \theta = \cup \quad (\text{form VIII})$$

(d) the transistor circuits of the eight forms I-VIII;

(e) the transistor circuits of the form IX, X;

(f) the cryotron circuits of the following forms

$$\omega = \perp, \quad \theta = \perp \quad (\text{form VII})$$

$$\omega = \cap, \quad \theta = \cap \quad (\text{form VIII})$$

D. In the classical case, the following simplification is made

$$\begin{aligned} xyz \cup \bar{x}yz \cup x\bar{y}z \cup xy\bar{z} \\ = xyz \cup \bar{x}yz \cup xyz \cup x\bar{y}z \cup xyz \cup xy\bar{z} \\ = (x \cup \bar{x}) yz \cup (y \cup \bar{y}) xz \cup (z \cup \bar{z}) xy \\ = yz \cup xz \cup xy \end{aligned} \quad (18)$$

by virtue of the idempotence law

$$z \cup z = z \quad (19)$$

To indulge in this type of computation, it is necessary to assume that

V. The following equality is true

$$\begin{aligned} \mathcal{A}_0 \omega \mathcal{A}_0 \omega \mathcal{A}_1 \omega \dots \omega \mathcal{A}_r \\ = \mathcal{A}_0 \omega \mathcal{A}_1 \omega \dots \omega \mathcal{A}_r \end{aligned}$$

This property is valid for the operations  $\cup, \cdot, \cap, \vee$ , but it is not valid for  $+$  and  $\sim$ .

E. It is known that in the classical case, there can be the following type of simplification

$$\begin{aligned} xy \cup \bar{y}z \cup xz &= xy \cup \bar{y}z \cup xyz \cup x\bar{y}z \\ &= (xy \cup xyz) \cup (\bar{y}z \cup x\bar{y}z) \\ &= xy \cup \bar{y}z \end{aligned} \quad (21)$$

The problems arising from this type of simplification constitute the originality of Quine's method.

A start is made with an expression such as eqn (13) where the  $z_{ih}$  have been replaced by  $L_{\alpha}(x_{\beta})$  as in the expressions provided by eqn (4).

An expression of the form

$$\mathcal{A} = L_{\alpha_1}(x_{a_1}) \theta \dots \theta L_{\alpha_r}(x_{a_r}) \quad (22)$$

is called a simple expression.

If  $\mathcal{A}$ ,  $\mathcal{D}$  are simple expressions

$$\mathcal{A} \propto \mathcal{D} \quad (23)$$

provided that

$$\{L_{\alpha_1}(x_{a_1}), \dots, L_{\alpha_r}(x_{a_r})\} \supset \{L_{\beta_1}(x_{b_1}), \dots, L_{\beta_s}(x_{b_s})\} \quad (24)$$

where the inclusion is considered in the sense of the set theory.

It is obvious that, on the basis of principles I-V, it can be deduced from eqn (23), that\*

$$\begin{aligned} L_{\alpha_1}(x_{a_1}) \theta \dots \theta L_{\alpha_r}(x_{a_r}) \theta L_{\beta_1}(x_{b_1}) \theta \dots \theta L_{\beta_s}(x_{b_s}) \\ = L_{\alpha_1}(x_{a_1}) \theta \dots \theta L_{\alpha_r}(x_{a_r}) \end{aligned} \quad (25)$$

Since eqn (23) is equivalent to eqn (24), it is easy to deduce that the relation  $\propto$  between the simple expressions is a relation of partial order, i.e.

$$\mathcal{A} \propto \mathcal{A}$$

$$\text{if } \mathcal{A} \propto \mathcal{D} \text{ and } \mathcal{D} \propto \mathcal{L}, \text{ then } \mathcal{A} \propto \mathcal{L}$$

$$\text{if } \mathcal{A} \propto \mathcal{D} \text{ and } \mathcal{D} \propto \mathcal{A}, \text{ then } \mathcal{A} = \mathcal{D}$$

The relation of contiguity will be defined between two simple expressions of the form

$$\begin{aligned} \mathcal{A} = L_{\alpha_1}(x_1) \theta \dots \theta L_{\alpha_{s_1-1}}(x_{s_1-1}) \theta L_{\alpha_{s_1+1}}(x_{s_1+1}) \theta \\ \dots \theta L_{\alpha_{s_t-1}}(x_{s_t-1}) \theta L_{\alpha_{s_t+1}}(x_{s_t+1}) \theta \dots \theta L_{\alpha_n}(x_n) \end{aligned} \quad (26)$$

$$\begin{aligned} \mathcal{D} = L_{\beta_1}(x_1) \theta \dots \theta L_{\beta_{s_1-2}}(x_{s_1-1}) \theta L_{\beta_{s_1+1}}(x_{s_1+1}) \theta \\ \dots L_{\beta_{s_t-1}} \theta (x_{s_t-1}) \theta L_{\beta_{s_t+1}}(x_{s_t+1}) \theta \dots \theta L_{\beta_n}(x_n) \end{aligned}$$

(the missing letters  $x_{s_1}, \dots, x_{s_t}$  are the same in  $\mathcal{A}$  and  $\mathcal{D}$ ) by

$$|\alpha_1 - \beta_1| + \dots + |\alpha_n - \beta_n| = 1 \quad (27)$$

F. It is proposed to simplify methodically the expression

$$\mathcal{E} = \mathcal{M}_1 \omega \dots \omega \mathcal{M}_t \quad (28)$$

where  $\mathcal{M}_i$  are simple expressions with  $n$  letters  $x$ .

If the simple expressions  $\mathcal{M}_i$  and  $\mathcal{M}_j$  are compared in case they are contiguous, if the letters  $x$  (third principle) are re-ordered, and if use is made of eqn (17) (fourth principle); a simple expression  $\mathcal{M}_{(ij)}$  is formed with  $n - 1$  letters. One thus obtains several simple expressions  $\mathcal{M}_{(ij)1}, \dots, \mathcal{M}_{(ij)t1}$  with  $n - 1$  letters. By virtue of the fifth principle

$$\mathcal{E} = \mathcal{M}_1 \omega \dots \omega \mathcal{M}_t \omega \mathcal{M}_{(ij)1} \omega \dots \omega \mathcal{M}_{(ij)t1}$$

\* It can be seen that this equality cannot be written as

$$\mathcal{A} \theta \mathcal{D} = \mathcal{A}$$

since  $\theta$  is not associative (in particular  $\top$  and  $\perp$  are not associative).

Compare the simple expressions  $\mathcal{M}_{(ij)1}$  and  $\mathcal{M}_{(ij)2}$  with  $n$  letters; if two of them are contiguous, simplify a letter according to the fourth principle and obtain expressions  $\mathcal{M}_{(ijhk)1}, \dots, \mathcal{M}_{(ijhk)t1}$  with  $n - 2$  letters. The operation is reiterated as many times as possible and it will yield  $\mathcal{E}$  in the following form:

$$\mathcal{E} = \mathcal{M}_1 \omega \dots \omega \mathcal{M}_{(ij)1} \omega \dots \omega \mathcal{M}_{(ij)t1} \omega \dots \omega \mathcal{M}_{(ijhk)1} \omega \dots \quad (29)$$

Among all these  $\mathcal{M}$ , the name of prime implicant is given to  $\mathcal{M}_\alpha$  if, from

$$\mathcal{M}_\alpha \propto \mathcal{M}_\beta$$

can be deduced  $\mathcal{M}_\alpha = \mathcal{M}_\beta$ . Let

$$B_1, \dots, B_v \quad (30)$$

be the totality of the prime implicants.

It is obvious that if, in the process of going from eqn (28) to (29), the fifth principle is not used, except when strictly necessary

$$\mathcal{E} = B_1 \omega \dots \omega B_v \quad (31)$$

but the example of eqn (29) shows that even this expression can be simplified to

$$\mathcal{E} = B_{\mu_1} \omega \dots \omega B_{\mu_v} \quad (32)$$

which contains only some of the prime implicants.

For this end perform again the simplification process described at the beginning of this paragraph. Starting from eqn (28) and considering each  $\mathcal{M}_i$ ; let  $B_{j1}, \dots, B_{js}$  be those prime implicants for which

$$\mathcal{M}_i \propto B_j$$

There is at least one among the  $B_j$  which satisfies this condition. Consider the propositions  $p_j$  = the prime implicant  $B_j$  appears in eqn (32). Since the simple expression  $\mathcal{M}_i$  appears in eqn (28), the proposition

$$q_i = p_{j1} v \dots v p_{js} \quad (33)$$

must be true.

Forming the proposition

$$P = q_1 \& \dots \& q_t \quad (34)$$

By expanding eqn (34)

$$P = R_1 v \dots v R_w \quad (35)$$

where  $R_i$  are propositions of the form

$$p_{\sigma_1} \& \dots \& p_{\sigma_t} \quad (36)$$

which means that

$$\mathcal{E} = B_{\sigma_1} \omega \dots \omega B_{\sigma_t} \quad (37)$$

However,  $B_{\sigma_1}, \dots, B_{\sigma_t}$  may be non-different, and therefore the application of the fifth principle will lead to a simplified form of  $\mathcal{E}$ .

Such a form corresponds at each  $R_i$ .

G. The simplification problem may be stated not only for the canonical forms

$$[L_{\alpha_1}(x_1) \theta \dots \theta L_{\alpha_n}(x_n)] \omega \dots \omega [L_{\alpha_r}(x_1) \theta \dots \theta L_{\alpha_n}(x_n)]$$

but also for other normal forms, such as

$$\begin{aligned} [L_{\alpha_{11}}(x_{\beta_{11}}) \theta \dots \theta L_{\alpha_{1p_1}}(x_{\beta_{1p_1}})] \omega \\ \dots \omega [L_{\alpha_{r1}}(x_{\beta_{r1}}) \theta \dots \theta L_{\alpha_{rpr}}(x_{\beta_{rpr}})] \end{aligned}$$

since, by applying eqn (4), the missing variables can be introduced into every simple expression.

Classic example

$$\begin{aligned}\mathcal{E} &= xy \cup \bar{y}z \cup xz = xyz \cup xy\bar{z} \cup \bar{y}z \cup xz \\ &= xyz \cup xy\bar{z} \cup x\bar{y}z \cup \bar{x}\bar{y}z \cup xz \\ &= xyz \cup xy\bar{z} \cup x\bar{y}z \cup \bar{x}\bar{y}z \cup xyz \cup x\bar{y}z \\ &= xyz \cup xy\bar{z} \cup x\bar{y}z \cup \bar{x}\bar{y}z\end{aligned}$$

The expression considered is of the form

$$\mathcal{E} = \mathcal{M}_1 \cup \mathcal{M}_2 \cup \mathcal{M}_3 \cup \mathcal{M}_4$$

with

$$\begin{aligned}\mathcal{M}_1 &= xyz \\ \mathcal{M}_2 &= xy\bar{z} \\ \mathcal{M}_3 &= x\bar{y}z \\ \mathcal{M}_4 &= \bar{x}\bar{y}z\end{aligned}$$

According to the fifth principle ( $\mathcal{M}_{1,2}$ ), ( $\mathcal{M}_{2,3}$ ), ( $\mathcal{M}_{3,4}$ ) are introduced

$$\begin{aligned}\mathcal{M}_{(12)} &= xy \\ \mathcal{M}_{(13)} &= xz \\ \mathcal{M}_{(34)} &= \bar{y}z\end{aligned}$$

The prime implicants are

$$\begin{aligned}B_1 &= \mathcal{M}_{(12)} \\ B_2 &= \mathcal{M}_{(13)} \\ B_3 &= \mathcal{M}_{(34)}\end{aligned}$$

$$\begin{aligned}\mathcal{M}_1 &\propto B_1 \\ \mathcal{M}_2 &\propto B_2 \\ \mathcal{M}_3 &\propto B_1 \\ \mathcal{M}_3 &\propto B_2 \\ \mathcal{M}_3 &\propto B_3 \\ \mathcal{M}_4 &\propto B_3\end{aligned}$$

$$\mathcal{E} = B_1 \cup B_2 \cup B_3 \text{ is eqn (29)}$$

New example

$$\begin{aligned}\mathcal{E}^* &= (x \perp y) \perp (\bar{y} \perp z) \perp (x \perp z) = (x \perp y \perp z) \\ &\quad \perp (x \perp y \perp \bar{z}) \perp (\bar{y} \perp z) \perp (x \perp z) \\ &= (x \perp y \perp z) \perp (x \perp y \perp \bar{z}) \perp (x \perp \bar{y} \perp z) \\ &\quad \perp (\bar{x} \perp \bar{y} \perp z) \perp (x \perp z) \\ &= (x \perp y \perp z) \perp (x \perp y \perp \bar{z}) \perp (x \perp \bar{y} \perp z) \perp (\bar{x} \perp \bar{y} \perp z) \\ &\quad \perp (x \perp y \perp z) \perp (x \perp \bar{y} \perp z) \\ &= (x \perp y \perp z) \perp (x \perp y \perp \bar{z}) \perp (x \perp \bar{y} \perp z) \perp (\bar{x} \perp \bar{y} \perp z)\end{aligned}$$

$$\mathcal{E}^* = \mathcal{M}_1^* \cup \mathcal{M}_2^* \cup \mathcal{M}_3^* \cup \mathcal{M}_4^*$$

$$\begin{aligned}\mathcal{M}_1^* &= x \perp y \perp z \\ \mathcal{M}_2^* &= x \perp y \perp \bar{z} \\ \mathcal{M}_3^* &= x \perp \bar{y} \perp z \\ \mathcal{M}_4^* &= \bar{x} \perp \bar{y} \perp z\end{aligned}$$

$$\begin{aligned}\mathcal{M}_{(12)}^* &= x \perp y \\ \mathcal{M}_{(13)}^* &= x \perp z \\ \mathcal{M}_{(34)}^* &= \bar{y} \perp z\end{aligned}$$

$$\begin{aligned}B_1^* &= \mathcal{M}_{(12)}^* \\ B_2^* &= \mathcal{M}_{(13)}^* \\ B_3^* &= \mathcal{M}_{(34)}^*\end{aligned}$$

$$\begin{aligned}\mathcal{M}_1^* &\propto B_1^* \\ \mathcal{M}_2^* &\propto B_2^* \\ \mathcal{M}_3^* &\propto B_1^* \\ \mathcal{M}_3^* &\propto B_2^* \\ \mathcal{M}_3^* &\propto B_3^* \\ \mathcal{M}_4^* &\propto B_3^*\end{aligned}$$

$$(p_1vp_2) \& p_1 \& (p_2vp_3) \& p_3 = p_1 \& p_3 \text{ is true}$$

$$\begin{aligned}\mathcal{E} &= B_1 \cup B_3 \\ &= xy \cup \bar{y}z\end{aligned}$$

$$\begin{aligned}\mathcal{E}^* &= B_1^* \perp B_3^* \\ &= (x \perp y) \perp (\bar{y} \perp z)\end{aligned}$$

A parallel is drawn between the classical example exposed at the beginning of section V and another similar one.

To sum up, in order to apply Quine's method, it is necessary that the first to the fifth principles be valid.

H. On other occasions, the author has drawn attention to the fact that multiplicative elements occur in circuits with contacts and relays. A few examples follow.

(a) In real operating conditions, the armature with contacts does not change suddenly from the attracted or the repulsed

position. There exists also an intermediate position, in which the normally open contacts as well as the normally closed ones are open (the 'break before make' relays) (Figure 1) or else the normally open contacts as well as the normally closed ones are closed (the 'make before break' relay) (Figure 2).

(b) The polarized relays whose armatures possess three possible positions, namely, resting, attraction and repulsion positions.

(c) The codified relays: examples (Figures 3, 4, 5) of codified

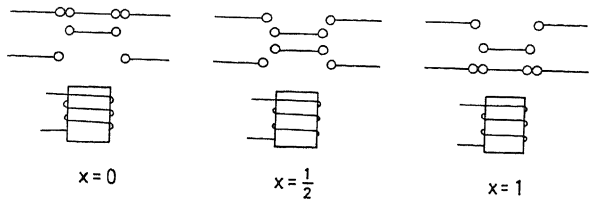


Figure 1

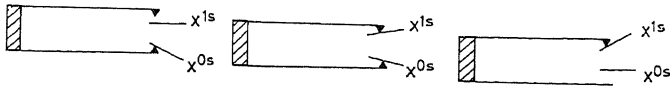


Figure 2

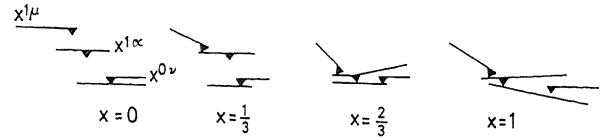
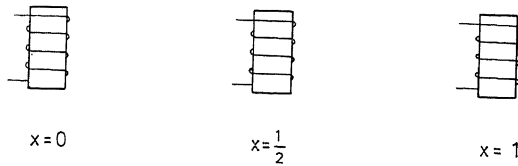


Figure 3

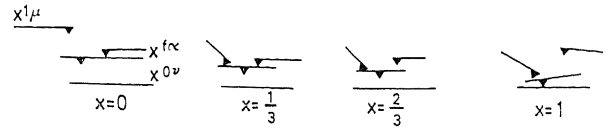


Figure 4

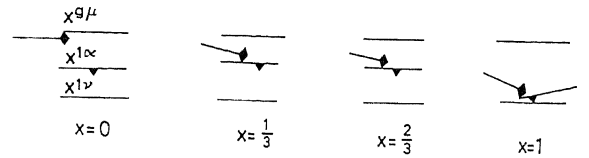


Figure 5

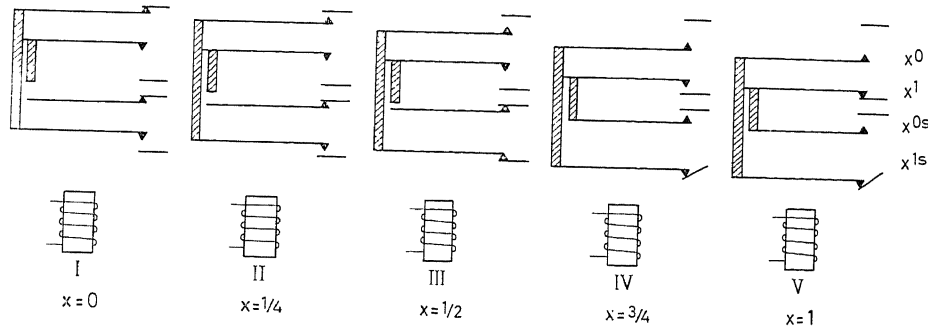


Figure 6

four-positional relays, taken from the book of Keister-Ritchie-Washburn, and the example given by Ivanin (Figure 6).

(d) The 'step by step' searcher or selector.

To each element a number of contacts can be associated, namely:

(i) In real operation of the relays  $X$  of the 'break before make' type, there exist normally open contacts  $\varphi_1^0(X)$  and normally closed contacts  $\varphi_0^0(X)$ ; in real operation of the 'make before break' relays  $X$  there are also normally open contacts  $\varphi_1^s(X)$  and normally closed contacts  $\varphi_0^s(X)$ .

(ii) The polarized relays  $X$  have contacts for the attraction position  $\varphi_1(X)$  and contacts for the repulsion position  $\varphi_2(X)$ ; some of the polarized relays also have, contacts for the resting position  $\varphi_0(X)$ , but some others, with an unstable neuter, lack such contacts.

(iii) The codified relays  $X$  possess several types of contacts.

(iv) The selector  $S$  with  $\nu$  steps has the brush contacts  $\varphi_0(S), \dots, \varphi_{\nu-1}(S)$ .

I. To each  $n$ -positional element, two sets of  $n$  elements are associated:

(a) The ring of residue classes modulo  $n$

$$\mathcal{J}/(n) = (0, 1, \dots, n-1) \quad (38)$$

with two operations: the addition and multiplication modulo  $n$  denoted by  $+$  and  $\cdot$ .

(b) The  $n$ -valent Lukasiewicz algebra

$$L_n = \left(0, \frac{1}{n-1}, \dots, \frac{n-2}{n-1}, 1\right) \quad (39)$$

with the natural order relation and with the operations

$$a \cup b = \max(a, b)$$

$$a \cap b = \min(a, b) \quad (40)$$

The Lagrange functions are denoted by  $L_\alpha(x)$

$$L_\alpha(\alpha) = 1$$

$$L_\alpha(\beta) = 0, \alpha \neq \beta \quad (41)$$

with  $\alpha \in \mathcal{J}/(n)$  respectively  $\alpha \in L_n$ . The dual functions  $\bar{L}_\alpha(x)$  are introduced with

$$\bar{L}_\alpha(\alpha) = 0$$

$$\bar{L}_\alpha(\beta) = 1, \alpha \neq \beta \quad (42)$$

There is a Lagrange interpolation formula in  $\mathcal{J}/(n)$

$$f(x_1, \dots, x_n) = \sum f(\alpha_1, \dots, \alpha_n) L_{\alpha_1}(x_1) \dots L_{\alpha_n}(x_n) \quad (43)$$



and two interpolation formulae in  $L_n$

$$f(x_1, \dots, x_n) = \bigcup_{\alpha} [f(\alpha_1, \dots, \alpha_n) \cap L_{\alpha_1}(x_1) \cap \dots \cap L_{\alpha_n}(x_n)] \quad (44)$$

$$f(x_1, \dots, x_n) = \bigcap_{\alpha} [f(\alpha_1, \dots, \alpha_n) \cup \bar{L}_{\alpha_1}(x_1) \cup \dots \cup \bar{L}_{\alpha_n}(x_n)] \quad (45)$$

$$\text{Giving} \quad [L_{\alpha}(x)]^2 = L_{\alpha}(x) \quad (46)$$

$$L_{\alpha}(x) L_{\beta}(x) = 0, \alpha \neq \beta \quad (47)$$

$$L_{\alpha}(x) \cap L_{\beta}(x) = 0, \alpha \neq \beta \quad (48)$$

$$L_1(x) + \dots + L_n(x) = 1 \quad (49)$$

$$L_1(x) \cup \dots \cup L_n(x) = 1 \quad (50)$$

to  $\omega$  and  $\theta$  in eqn (13), the following values can be given

	$\omega$	$\theta$	
XI	+	·	
XII	∪	∩	
XIII	∩	∪	(51)

In  $\mathcal{J}(n)$  respectively in  $L_n$ , principles I, II, III, and V are satisfied for the substitutions XI, XII, XIII of  $\omega$ ,  $\theta$ .

To the fourth principle, must be substituted

IV\*. The following equality is true

$$\begin{aligned} & [L_0(z) \theta y_1 \theta \dots \theta y_r] \omega [L_1(z) \theta y_1 \theta \dots \theta y_r] \omega \\ & \dots \omega [L_{n-1}(z) \theta y_1 \theta \dots \theta y_r] \omega t_1 \omega \dots \omega t_s \\ & = (y_1 \theta \dots \theta y_r) \omega t_1 \omega \dots \omega t_s \end{aligned}$$

Principles I, II, III, IV\*, V allow application of Quine's method to multipositional elements.

In order to have a better understanding of the method one should introduce also the notions of formula of structure, function of work and functional equivalence in these general cases.

## Bibliography

The author has published the following volumes

- <sup>1</sup> *Teoria algebrica a mecanismelor automate* (Algebraic theory of switching circuits). 1959; București; Editura tehnică
- <sup>2</sup> *Scheme cu comanda directă cu contacte și relee* (Combinational switching circuits with contacts and relays). 1959. București; Ed. Acad. Rep. populare Romine.
- <sup>3</sup> *Functionarea în mai mulți timpi a schemelor cu relee ideale* (Sequential ideal operation of relay switching circuits). București, Ed. Acad. Rep. populare Romine.
- <sup>4</sup> *Circuite cu transistori* (Transistor switching circuits). 2 Vols. 1962. Ed. Acad. Rep. populare Romine.  
The interpolation formulae have been taken from
- <sup>5</sup> *Simplificarea circuitelor cu tuburi electronice, cu transistori și criotroni* (Simplifying the circuits containing electronic tubes, transistors and cryotrons). *Revue Math. pures et appl.* (1959) 497  
For the electronic tubes circuits see 5 and the authors work
- <sup>6</sup> Algebraic theory of circuits with electronic tubes, Physico-mathematical Inst. Bulgarian Acad. Science 4 (37) (1961) p. 7  
For circuits with transistors see 4, 5, 6 and the author's work

- <sup>7</sup> Sur la théorie algébrique des circuits logiques à transistors. *Automatisme*, 7 (1962) 136  
For the cryotron circuits, see 5  
For the real operating of circuits with ordinary switching relays, see 1 and Gh. Ioanin's work
- <sup>8</sup> A supra teoriei algebrice a contactelor multipozitionale și aplicațiile ei la studiul contactelor reale (On the algebraic theory of multi-positional contacts and its application to the study of real contacts). *Bul. sti. Acad. Repub. rom. sec. sti. mat. fiz.* 7 (1955) 231.  
See also the author's work with Ioanin
- <sup>9</sup> A supra funcționării schemelor cu butoni reali (On the operation of circuits with real buttons). *Bul. sti. Acad. Repub. Pop. rom. sec. sti. mat. fiz.* 7 (1955) 33, see also 1
- <sup>10</sup> Sur l'application des logiques à trois valeurs à l'étude des schémas à contacts et relais, *Act. Congr. int. Automat., Paris*, 14-24 June 1956, p. 48
- <sup>11</sup> Aplicațiile logice trivalente în studiul funcționării reale a schemelor cu contacte și relee (The applications of trivalent logic to the study of real operating of circuits with contacts and relays). *Bul. matem. Soc. sti. mat. fiz. Repub. Pop. rom.* 1 (49) (1957) 197
- <sup>12</sup> Sinteza schemelor cu contacte și relee în funcționare reală (Synthesis of circuits with contacts and relays, under real operating conditions). *Bull. matem. Soc. sci. math. phys. Roum.* 3 (51) (1959) 65  
A volume on the real operating conditions is in the press  
For circuits with polarized relays see 1 and the author's works
- <sup>13</sup> Sur la synthèse des schémas à relais polarisés. Bulgarian Academy of Science 2 (1957) 121
- <sup>14</sup> Sur la théorie algébrique des mécanisme automatiques. Synthèse des schémas à relais polarisés. *Ber. int. matem. Koll., Dresden*, 22-27 November 1955, Aktuelle Probleme der Rechentechnik. Deutscher Verlag der Wissenschaft. 1957. Berlin  
For circuits with codified relays see 1 and the author's works
- <sup>15</sup> Logica matematică și tehnica modernă. Logicele cu mai multe valori și circuitele cu contacte și relee (Mathematical logic and modern technique. The logics with several values and the circuits with contacts and relays). *Probleme filosofice ale științelor naturii*. 1960. ISRS Acad. Rep. populare Romine  
For circuits with selectors, see 1 and Gh. Ioanin's works
- <sup>16</sup> Sinteza schemelor în care intră selectori (Synthesis of circuits with selectors). *Bul. sti. Acad. Repub. Pop. rom., ser. mat. fiz.* 8 (1956) 489; *Automat. Telemech.*, Moscow 19 (1958) 855
- <sup>17</sup> Sur un type de problème concernant les schémas à sélecteurs. *Acta Logica, Bucharest* (1958) 187

## Further Work

The theory of abstract finite automata has been reduced to the two equations:

$$X_{N+1} = F(K_N, X_N) \quad (52)$$

$$W_N = G(K_N, X_N) \quad (53)$$

where the domains of  $K$ ,  $W$ ,  $X$  are the sets of inputs, outputs, and internal states of the automaton.

An abstract description of the real automata can be made as follows: Suppose the automaton is built with pushbuttons, ordinary relays and bulbs.

The state  $K$  of the pushbuttons (which may be operated or released) is determining the state  $H$  of the contacts (open or closed) of these buttons:

$$H_N = \Phi(K_N) \quad (54)$$

For each relay we shall consider three variables: the variable  $x$  associated to the operation state of the armature, which can be

attracted or released; the variable  $y$  associated with the current in the relay coil (current or no current) and the variables  $z^0$  and  $z^1$  associated with the contacts which may be normally open or normally closed. Let  $X, Y, Z$  be the variables for the sets of all relays. Then the next internal state is determined by the former and the current in the former time interval.

$$X_{N+1} = \psi(X_N, Y_N) \quad (55)$$

The open or closed state of the contacts is due to the position of the armature

$$Z_N = \Lambda(X_N) \quad (56)$$

Eqns (55) and (56) are characteristic for each secondary element;

ordinary relay, polarized relay, latching relay, rotary switch, etc.

The structure of the network yields:

$$Y_n = \Omega(H_N, X_N) \quad (57)$$

The elimination of  $H, Y, Z$  gives eqn (52). However, the set of eqns (54), (55), (56) and (57) permits a better description of the automaton.

The simplification problem can be formulated in this theory; it concerns eqn (57) where the function  $\Omega$  is understood as a function generated by an expression which is the structure formula.

This problem is solved in the paper.

# Adaptive Control for a System with a Finite Number of States

S. PASZKOWSKI

## Summary

In this work a system with a finite number of states is considered. The dynamics of the system is unknown. On the basis of the given criterion function it is necessary to control the medium in such a way that a minimum of the losses expected is obtained.

It was proved that under these conditions for a certain medium, the controlling solutions must minimize the losses only over the current step. When the probabilities for the transition of states of the medium are unknown the solutions are adopted on the basis of an assembled experiment. The results of experiments are intended for the determination of the actual parameters of the medium. For the purpose of this determination the method of 'reliable intervals' was used. The algorithm for the adoption of controlling solutions is determined and the results are presented for the operation of automatic control for two types of media with unknown parameters. The results of experiments show that the algorithm used leads rapidly to the optimum range of controlling solutions.

## Sommaire

Dans ce rapport un système avec un nombre fini d'états est pris en considération. La dynamique de ce système n'est pas connue. En se basant sur un critère de performance, il s'agit d'ajuster ce système de façon à minimiser les pertes.

Il est prouvé que dans ces conditions et pour un certain système, il suffit de minimiser les pertes pour l'échelon courant. Quand la probabilité du passage d'un état à un autre n'est pas connue, les solutions sont obtenues expérimentalement par superposition. Ces résultats sont utilisés pour déterminer les paramètres du système. Dans ce but, on utilise la méthode des "intervalles fiables" ("reliable intervals"). L'algorithme conduisant à la solution est indiqué et les résultats sont présentés pour l'optimisation de deux types de systèmes avec paramètres inconnus. Les résultats expérimentaux montrent que l'algorithme utilisé conduit rapidement vers l'optimisation désirée.

## Zusammenfassung

Diese Arbeit betrachtet ein System mit einer endlichen Anzahl von Zuständen, dessen dynamisches Verhalten unbekannt ist. Auf Grund der gegebenen Gütefunktionen muß man das System so regeln, daß die zu erwartenden Verluste zu einem Minimum werden.

Wie bewiesen wurde, erfordern diese Bedingungen lediglich, daß die Verluste nur während des gegenwärtigen Schrittes minimal werden. Sind die Wahrscheinlichkeiten für den Übergang von einem Zustand des Systems auf den anderen unbekannt, so gewinnt man die Lösungen auf Grund von Beobachtungen am System, deren Ergebnisse zur Bestimmung der tatsächlichen Parameter des Systems dienen. Für diese Bestimmung wurde die Methode der „zuverlässigen Intervalle“ (reliable intervals) benutzt. Der Algorithmus, der zur Lösung der Regelung führt, wird bestimmt und die Ergebnisse für den Regelablauf von zwei Arten von Systemen mit unbekannten Parametern vorgelegt. Die Versuchsergebnisse zeigen, daß der benutzte Algorithmus schnell zum optimalen Bereich der Lösungen für die Regelung führt.

## Introduction

In this article a system with incomplete information concerning the medium is considered. Problems of this kind are encountered

in engineering, economics, and in systems of mass maintenance. In the systems of control with incomplete information regarding the behaviour of the medium, irregular and inaccurate controlling solutions may be adopted, which results in great losses. In connection with this the development of such an algorithm for controlling solutions, which would rapidly reduce the number of inaccurate and costly solutions, represents an important problem. The system, which will realize this algorithm, may be called the automatic system of control.

## Formulation of Problem

Let  $U = (u_1, u_2, \dots, u_z)$  be the set of actions at our disposal. These actions can occur only at discrete moments of time. On each step only one action,  $u(n) \in U$ , can occur.

Let  $X^* = (x_1^*, x_2^*, \dots, x_z^*)$  be the set of events, which can be received by receivers  $R$ . The event occurring on the  $n$ th step will be denoted by  $x^*(n)$ .

Let  $X = (x_1, x_2, \dots, x_z)$  be the set of events, which can be received by receivers  $L$ . The event occurring on the  $n$ th step will be denoted by  $x(n)$ .

In addition there is the criterion function  $S(x^*(n), x(n+1))$  which determines for each step the occurring events. This function is represented in the form of Table 1.

From Table 1 it follows that events  $x^*(n) = x(n+1)$  are the events desired. For any other event the 'penalty' represented by number  $r$  must be paid.

Event  $x^*$  may be regarded as the request received on the given step, and event  $x$ , as the realization of that request. When the realization is identical with the request there are no losses. Otherwise losses  $r$  occur.

Receivers  $R$  receive events  $x^*$  from medium  $A$ . The processes in medium  $A$  which have an effect on received events  $x^*$  may be described only in the form of a probability. In this

Table 1

$x^*(n) \backslash x(n+1)$	$x_1$	$x_2$	...	$x_z$
$x_1^*$	0	$r$		$r$
$x_2^*$	$r$	0		$r$
.	.			.
.	.			.
.	.			.
.	.			.
$x_z^*$	$r$	$r$	...	0

particular problem it was assumed that  $P(x^* = x_i^*) = p^* = 1/z$  where  $i = 1, 2, 3, \dots, z$ . This means that at each stage any event  $x^* \in X^*$  may occur with an equal probability.

Receivers  $L$  receive events  $x$  from medium  $B$ . The processes of medium  $B$  may be influenced by permissible actions  $u \in U$ . Nothing is known about the mechanism of the effect of processes of medium  $B$  and of adopted action on events  $x$ , except that such an effect does exist. The structure of the mathematical model which will be used for the finding of a connection between the adopted action and the received event  $x(n)$  is represented by the matrix for the probabilities of transition

$$(p_{ij}^k) \quad i, j, k = 1, 2, 3, \dots, z \quad (1)$$

where  $p_{ij}^k = P(x_i(n) \rightarrow x_j(n+1))$  is the probability of occurrence of the event  $x(n+1) = x_j$  when  $x(n) = x_i$ ;  $u(n) = u_k$ . In the problem considered  $p_{ij}^k$  are slowly changing unknown numbers.

The aim is that, under the above-defined conditions, the losses obtained should be at a minimum. This is the general aim of action of an organized system. In this system a stochastic process takes place and, therefore, the mentioned aim should be regarded as the realization of a minimum of mean expected losses. In connection with this the problem of automatic control is to produce on each step such actions for which the mean expected losses will be at a minimum.

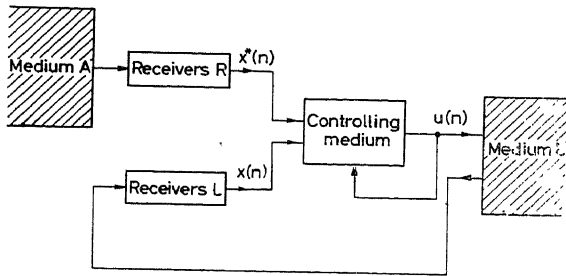


Figure 1

The block diagram of an organized system is shown in Figure 1. It may be assumed that in the given system there exists a series of actions, which solves the basic problem. This series of actions, for which a minimum of mean expected losses is obtained, will be called the 'decisive (determinative) strategy'. In the given system there may be several decisive strategies. With the change in the operating conditions of the system there is a change in the decisive strategy.

The information regarding the behaviour of the medium is incomplete in the system. Therefore it is impossible to determine directly the decisive strategy. In connection with this a new, additional problem for the automatic control is created. It is, thus, necessary to control the actions in such a way that the decisive strategy is obtained with the minimum of additional losses. This additional problem is reduced to the finding of the decisive strategy.

#### The Solution of the Problem for Known Probabilities $p_{ij}^k$

First of all, automatic control for the known probabilities  $p_{ij}^k$  will be considered. For this, the algorithm for the working out of the decisive strategy will be determined.

The mean expected losses will be determined as the losses on the  $N$ th step of the path, where  $N$  can be as large as desired.

For the determination of the corresponding algorithm the method of dynamic programming will be used. First, the losses over one step will be calculated for the following conditions.

(a) The initial condition for the approaching step is known.

$$x^*(n) = x_i^*$$

$$x(n) = x_i$$

Under these conditions the mathematical expectation for losses over a single step is determined by the formula:

$$v(x_i, x_i^*, u_k) = \sum_{j=1}^z p_{ij}^k S(x_i^*; x_j(n+1)) = (1 - p_{ii}^k) \cdot r \quad (2)$$

(b) The initial condition is given in the form:

$$x(n) = x_i$$

$p_i^* = \frac{1}{z}$  is the probability of occurrence of any  $x_i^* \in X^*$ . Under these conditions the mathematical expectation for the losses over a given step is determined by the formula:

$$v^*(x_i, u_k) = \sum_{l=1}^z p_l^* (1 - p_{il}^k) \cdot r = \frac{z-1}{z} \cdot r \quad (3)$$

From (3) it is seen that for the unknown initial condition, the mathematical expectation for losses over a step does not depend on the probabilities of transition  $p_{ij}^k$ .

On the basis of the method of dynamic programming it is possible to write down the following equation:

$$V_{n+N}(x_i(n), x_i^*(n)) = \min_{u_k \in U} \left\{ v(x_i, x_i^*, u_k) + \sum_{h=1}^z p_h^* \sum_{j=1}^z p_{ij}^k V_{n+N-1}(x_j(n+1), x_h^*(n+1)) \right\} \quad (4)$$

In this equation the events  $x^*(n+1)$ ,  $x^*(n+2)$ , ...,  $x^*(n+N)$  are given only in the form of probability  $p^* = \frac{1}{z}$ . Therefore, the second portion of the right-hand side of eqn (4) may be represented in the form:

$$\sum_{h=1}^z p_h^* \sum_{j=1}^z V_{n+N-1}(x_j(n+1), x_h^*(n+1)) = \frac{z-1}{z} (N-1) r \quad (5)$$

Under these conditions

$$V_{n+N}(x_i(n), x_i^*(n)) = \min_{u_k \in U} \left\{ (1 - p_{ii}^k) \cdot r + \frac{z-1}{z} (N-1) r \right\} \quad (6)$$

From this it is evident that the optimum solution is that solution which gives the minimum value for the mathematical expectation of losses over a step and this is clear intuitively. For the same probability of occurrence of event  $x_i^* \in U$  over each step, there is no point in planning the actions over  $N$  steps.

On the basis of the given reasoning, the following algorithm for the operation of the decisive strategy is adopted. For each step such actions are adopted, for which the mathematical expectation for step losses is at a minimum.

### Solution of Problem for Unknown Values of $p_{ij}^k$

For unknown probabilities  $p_{ij}^k$  it is impossible to apply the algorithm determined above. It is necessary to develop a new algorithm in order to find a decisive strategy. One of the methods for finding this is to use the information regarding the medium obtained during the time of operation of the system, and the gradual approach to the unknown strategy.

The results, obtained during the time of operation of automatic control, will be represented in the form:

$$\left( \frac{v_{ij}^k}{m_j^k} \right) \quad i, j, k = 1, 2, 3, \dots, z \quad (7)$$

where  $m_i^k$  is the number of adopted actions  $u_k$  for the initial condition  $x_i$ , and  $v_{ij}^k$  is the number of observed transitions from  $x_i(n)$  to  $x_j(n+1)$ , for  $m_i^k$  experiments.

These results will be used for the determination of unknown probabilities  $p_{ij}^k$ . The determination of the unknown values of  $p_{ij}^k$  will be made by means of confidence intervals. For each value of  $v_{ij}^k/m_i^k$  it is possible to calculate the confidence interval  $(P_{ijH}^k; P_{ijB}^k)$ , where  $P_{ijH}^k$  is the lower limit, and  $P_{ijB}^k$  the upper limit, of the interval.

The limits of intervals may be calculated from known expressions or they can be obtained from tables<sup>2</sup>.

The confidence interval determines the set of the hypothetically possible actual values of  $p_{ij}^k$ . With a high degree of reliability it can be assumed that the actual value of probability  $p_{ij}^k$  will be found in the above-defined interval.

Let the initial condition for the approaching step be:

$$\left. \begin{aligned} x^*(n) &= x_i^* \\ x(n) &= x_i \\ \frac{v_{il}^k}{m_i^k} &\rightarrow (P_{iil}^k, P_{iilB}^k) \end{aligned} \right\} \quad (8)$$

The working out of the decisive strategy represents the general problem of the system which consists in the control of actions. From this it follows that for a given initial condition it is necessary to choose action  $u_k \in U$  for which  $p_{ii}^k$  is at a maximum. However, since one knows only the confidence intervals, it is not possible to make a direct choice. In connection with this the following algorithm for the choice of action  $u_k \in U$ , is adopted: to choose such  $u_k \in U$  actions for which, for a given initial condition, there is a hope that probability  $p_{ii}^k$  has the maximum

value. This is identical with the method based on the choice of an action, for which there is a hope that the expected losses over a single step will be at a minimum.

It should be pointed out that the upper limit of the confidence interval  $P_{iilB}^k$  where  $k = 1, 2, 3, \dots, z$  represents the basis for the choice of the action. From this it follows that it is necessary to choose such values of  $u_k$ , for which the upper limit of the confidence interval has the maximum value. The result of the action will either confirm the correctness of choice or, in the case of a negative result, decrease the upper limit of the interval, which in the following intervals gives the possibility for the choice of another value for  $u_k$ . This method guarantees a sufficiently quick convergence of the actions being chosen towards the decisive strategy.

*Example 1*—In the given example the set of events  $X^*$  consists of three events ( $x_1^*$ ,  $x_2^*$ ,  $x_3^*$ ). Medium  $B$  is described by means of graphs, shown in Figure 2. The results of actions of

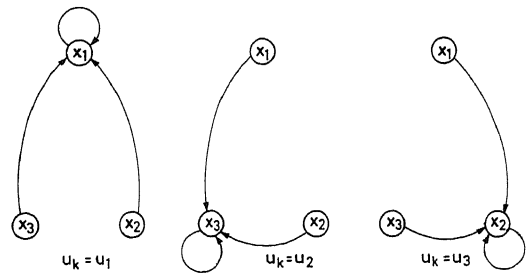


Figure 2

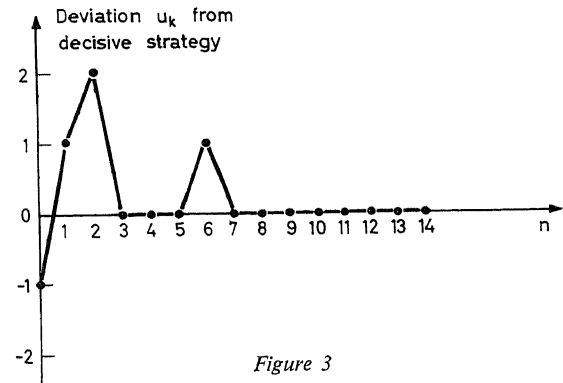


Figure 3

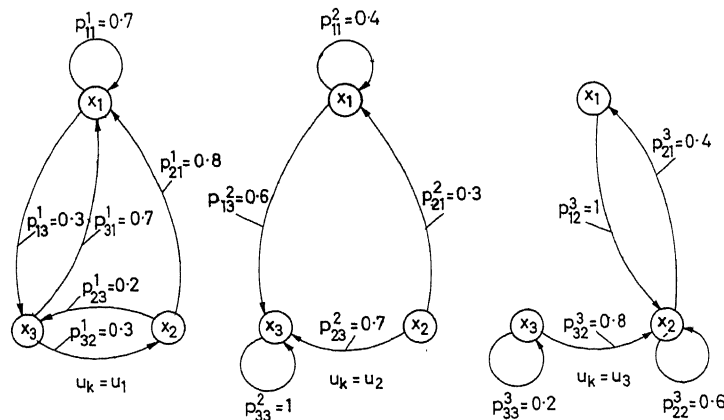


Figure 4

the system are shown on the graph (Figure 3). On this graph the deviations of the actual actions from the decisive strategy are seen. This system was investigated.

The results obtained indicate a rapid convergence of the actions towards the decisive strategy.

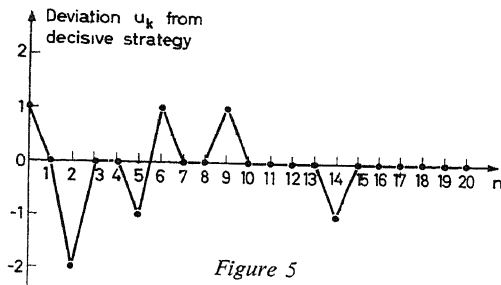


Figure 5

## DISCUSSION

J. SKLANSKY, R.C.A. Laboratories, Princeton, New Jersey, U.S.A.

The optimal decision strategies discussed in the paper are optimal with respect to performance, but not necessarily with respect to construction and maintenance costs. In fact, many so-called 'optimal' decision strategies can only be realized by digital computers. This situation is especially severe in the case where the transition probabilities are unknown, since the 'reliable intervals'  $p$ , i.e., the range of likely values of  $p_{ij}^k$ , must be recalculated at every step.

To overcome this shortcoming, the following procedure is suggested: Choose simple, easily constructed, easily maintained decision strategies, then determine the class of plant-environment combinations for which each of these decision strategies can be expected to yield good performance, taking into account the designer's *a priori* ignorance of the parameters of the plant and its environment.

This procedure has been applied to a simple problem in adaptive signal detection<sup>1</sup>, illustrated in Figure A. A binary information source

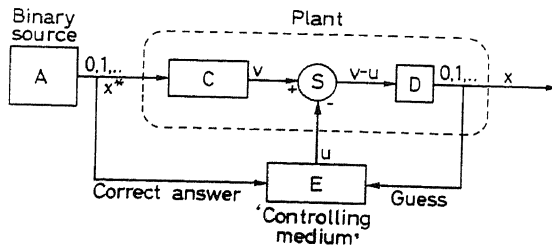


Figure A

A (a combination of the paper's 'medium A' and 'receivers R') emits  $x^*(n)$ , where the set of events  $X^*$  consists of just two events: 0 and 1. Wherever possible, the mathematical notation here is the same as in the paper. The plant ('medium B') consists of a noisy channel C, a subtractor S, and a sign detector D. The decision strategy is incorporated in the 'controlling medium' E. The decision strategy is very simple: when the plant's guess

$$u(n+1) = \begin{cases} u(n) + 1 & \text{if } u(n) < 3 \\ 3 & \text{if } u(n) = 3 \end{cases}$$

When  $x(n+1) = 0$  and  $x^*(n) = 1$ , then

$$u(n+1) = \begin{cases} u(n) - 1 & \text{if } u(n) > 1 \\ 1 & \text{if } u(n) = 1 \end{cases}$$

*Example 2*—In this example the behaviour of medium B has changed. Medium B is represented by the graph of Figure 4. The results obtained are shown in Figure 5. In this case also, a rapid convergence towards the desired strategy is obtained.

## References

- 1 FELDBAUM, A. A. Information storage in closed systems of automatic control. *Izv. Akad. Nauk SSSR, OTN, Energ. i Avt.* 4 (1961)
- 2 YANKO, YA. *Mathematical-statistical Tables*, 1961
- 3 BELLMAN, R. *Dynamic Programming*. 1957. New York
- 4 BUSK, R. and MOSTELLER, F. *Stochastic Models for Learning*. 1955

This is called 'simple incremental' decision strategy. The resulting Markov chain model of the closed-loop process is specified by the following stochastic matrix:

	1	2	3
1	$1 - \frac{\rho}{2}(1 + \alpha)$	$\frac{\rho}{2}(1 + \alpha)$	0
2	$\frac{1 - \rho}{2}(1 - \alpha)$	$\frac{1}{2}(1 + \alpha)$	$\frac{\rho}{2}(1 - \alpha)$
3	0	$\frac{1 - \rho}{2}(1 + \alpha)$	$1 - \frac{1 - \rho}{2}(1 + \alpha)$

In this matrix the states are the values of the control actions  $u$ . The variable  $\rho$  is a parameter of the information source, and  $\alpha$  is a parameter of the plant.

Suppose the performance index, denoted by  $z(n)$ , is defined as probability of a correct guess. Symbolically,

$$z(n) = \Pr[x(n) = x^*(n-1)]$$

Suppose 'good' performance is defined as  $z(\infty) \geq 0.8$ . Using these definitions one may compare the effectiveness of the simple incremental strategy (the 'closed-loop' operation) with an 'open-loop' operation in which  $u(n) \equiv 2$  for all  $n$ . To do this the regions in the  $\rho\alpha$  plane corresponding to  $z(\infty) \geq 0.8$  for the open-loop and closed-loop operations were computed. The results are shown in Figure B. The 'good' regions in the  $\rho\alpha$  plane, corresponding to  $z(\infty) \geq 0.8$  are shown shaded. Note that the closed-loop operation yields a substantially larger 'good' region than that of the open-loop operation.

When the transition<sup>1</sup> probabilities of the plant are unknown, it is possible, as indicated in the paper, to estimate 'reliable intervals' for an 'ignorance rectangle' in the parameter space, as indicated by the dotted lines in Figure B. It is suggested, that the designer should have at his disposal an 'atlas' of performance contours similar to those in Figure B. The designer will then be able to choose the least expensive decision strategy whose 'good' region covers—or almost covers—the ignorance rectangle of his particular problem.

This approach to the design of decision strategies promises to reduce substantially the construction and maintenance costs for a

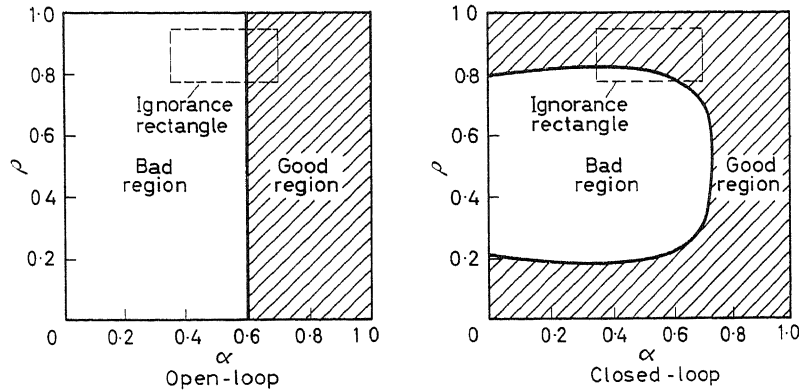


Figure B

given performance requirement, taking into account the *a priori* uncertainties of the environment and the unreliability of the plant.

#### Reference

- <sup>1</sup> SKLANSKY, J. *A Markov chain model of adaptive signal detection*, presented at the 1963 Bionics Symposium, March 19–21, 1963, Dayton, Ohio, U.S.A.

S. PASZKOWSKI, *in reply*

Mr. Sklansky has proposed a very interesting approach to the solution which I investigated in the present paper. I agree that the development of rational solving strategies and the analysis of the performance of a system controlled by this solving strategy is of great importance. I have proposed one of the possible algorithms for developing a solving strategy.

V. W. EVELEIGH, *General Electric Co., Syracuse, N.Y., U.S.A.*

The author presents several interesting examples involving straightforward application of probability and dynamic programming principles. Although the applications are perhaps unique, the principles upon which they are based are well known.

I wonder if the author has considered, in addition to his single dimensional example with first-order memory (Example 2), any problems with higher-order memory in which the transition probabilities depend upon present system state and also on the previous state and perhaps others prior to that? This is equivalent to dealing with more elements in the state vector and considering only first-order memory. This should be an interesting problem. Also, when the transition probabilities are unknown and must be measured as the solution proceeds, there is a potential solution instability problem. The probability values may be changing more rapidly than the system is able to follow using only the information obtained from relatively infrequent sampling of response. This point should be mentioned in the paper. If the rate of change is very slow, convergence is assured, assuming, of course, that the measurements of state are reliable. Finally, the significance of Figures 3 and 5 is not completely clear to me from the text. I would appreciate a further comment on their meaning by the author.

S. PASZKOWSKI, *in reply*

I should like to point out that in the studied problem medium  $B$  has memory, i.e., the probability for the occurrence of a definite event  $x(n+1)$  depends on the state  $x(n)$ .

An interesting remark by Mr. Eveleigh is the possibility of instability. I have worked on this problem and the results obtained are concerned with the dependence of the quantity of memory on the rate of variation of transform probabilities.

Concerning the second part of the question, it must be said that in Figures 3 and 5 differences between the number of solution, which at the same time must be also the number of the realized solution, are plotted.

CHUAN-SHAN WANG, *Institute of Automation, Academy of Sciences, Peking, China*

I should like to ask Professor Paszkowski the following questions:

- (1) What is the practical significance of his paper?
- (2) Has the author constructed a practical model of such automata, and what are his experimental results?
- (3) How did the author compute the results given in Figures 3 and 5 of his paper?
- (4) In my opinion, problems with unknown values of  $p_{ij}^k$  can be solved by known principles of modern statistics theory of communication. When the *a priori* probability is uniformly distributed it will yield the same results as obtained in this paper. What is the author's opinion?

S. PASZKOWSKI, *in reply*

(1) The algorithm in question can be applied to control such complex systems as transport systems and others in the case when the probability of the transition from one state to another is unknown.

(2) An automaton operating according to a given algorithm has a form of programme for a digital computer. This computer was working with an experimental object. The results obtained confirm the rapid convergence towards the solving strategy.

(3) The results presented in Figures 3 and 5 were obtained by means of a computer.

(4) Unknown probabilities  $p_{ij}^k$  can be estimated directly from the uniform distribution of *a priori* probabilities, but it is necessary to improve these estimates by means of experiments. The method of uniform distribution of probabilities *a priori* gives no possibility of controlling such experiments in order to obtain the improved estimates of necessary probabilities.

# THEORY OF OPTIMAL SYSTEMS

## The Synthesis of Optimal Regulators

A survey by A. M. LETOV

### Introduction

In a short report one cannot describe the content of the huge flow of literature that has been devoted to optimal control. The compilation of a survey has been made still more difficult by the fact that eminent mathematicians of many countries are taking part in the investigation of optimal control processes.

Outstanding successes have been achieved in their efforts to make the reasoning schematic and as general as possible, and to subordinate it to comprehensive laws of mathematical formalism. The methods of the maximum principle and of dynamic programming serve as a brilliant illustration of this. Thus any attempt to revise the theory of optimal control will inevitably verge on an assessment of the established mathematical canons, which the author dares not undertake.

These difficulties become insuperable when one tries to shed light on the historical aspects of the theory, and these will not be considered.

Hence, the author decided simply to set out his own views on the contemporary state of the problems in synthesis of optimal regulators for continuous systems.

In forming these views, the author has not relied only on reading the literature, but has sought support in the most instructive discussions he has had with Academician L. S. Pontryagin, Professor A. I. Lurye (corresponding member of the U.S.S.R. Academy of Sciences), Professor Ye. A. Barbashin, Professor V. I. Zubov, and Professor N. N. Krasovskiy.

### Setting of the Fundamental Problem

#### Programming of Brachistochronic Motions

There are two classes of optimal control problems. The first includes problems in programming optimal motions which have extremal properties that are specified in advance. The programming problem for deterministic systems is formulated as follows.

The differential equations for the controlled plant are given:

$$\dot{x} = X(x, \mu, t) \quad (1)$$

Here  $x \{x_1, \dots, x_n\}$  is the state vector,  $\mu$  is a scalar control function\*,

\* For the sake of simplicity one considers only plants with a single actuator. The following discussions can be generalized without any difficulty in principle to plants with a number of actuators.

$X \{X_1, \dots, X_n\}$  is the generalized force vector, defined over the region  $N(x, \mu) \geq 0$  characterized by the inequality

$$\begin{aligned} R(x, \mu, t) &\geq 0 \\ R \{R_1, \dots, R_m\} \end{aligned} \quad (2)$$

The system (1), (2) is defined for  $t \in [t_i, t_f]$ .

The boundary conditions are given, which are expressed symbolically as

$$(i, f) = 0 \quad (3)$$

A certain class of comparison functions is chosen (e.g.  $x$  the class of smoothed-segment functions and  $\mu$  the class of continuous-segment functions) and one determines for this the optimizing functional

$$\Phi = \tilde{\Phi}(x, \mu, t) \quad (4)$$

Functions  $x$  and  $\mu$  satisfying (1), (2) and (3) are called *permissible*.

The programming problem is to find from the permissible  $x$  and  $\mu$  those functions which make the functional (4) a minimum.

It is assumed that  $X, R, \Phi$ , and also conditions (3) are such that the required solution exists uniquely and has the form:

$$x = x^*[t, (i, f)], \quad \mu = \mu^*[t, (i, f)] \quad (5)$$

The motions (5) can be called brachistochronic, since they are expressed in parametric form, i.e. just as the solution of the famous problem propounded by Bernoulli was 267 years ago. It is material to emphasize that the programming problem is a transferred problem of the particle dynamics of the controlled plant, where the forces and motions are partly given and it is required to find those that are missing.

As such, *this problem lies outside the field of the direct scientific interests of specialists in automatic control*. In fact, to set up a mathematical model of the problem (to know the forces) requires a profound knowledge of the physical laws to which the plant (1) is subjected, and a precise definition of the ultimate aims in controlling it. The latter is the prerogative of scientists and engineers in the field of applied science—particle dynamics—to which the plant belongs. It is they who are capable of evaluating the extent to which the solution (5) actually meets the objective assigned.



The automation expert is called in to indicate the method and means for actually achieving the programmed motion, and handles it as a premise. Of course he does not lack interest in the model itself, in its properties and the peculiarities of solution (5).

The focus of his scientific interest lies, however, in the different problem of optimal control.

### Synthesis of Optimal Regulators. Basic Problem

The problem begins as follows. Let the solution (5) be called an unperturbed motion in Liapunov's sense. It is easily verified that solution (5) cannot be achieved, if only because state  $i^*$  (3) cannot be precisely realized. Hence the actual motion will be described by the functions

$$x = x^* + \eta, \quad \mu = \mu^* + \xi \quad (6)$$

where  $\eta \{ \eta_1, \dots, \eta_n \}$ ,  $\xi$  is a Liapunov perturbed motion, while  $\eta(0)$ ,  $\xi(0)$  is the initial perturbation\*. The perturbed motion satisfies the equations

$$\dot{\eta} = B\eta + m\xi + \Xi(\eta, \xi, t) \quad (7)$$

The right-hand side of (7) are some form of series expansions of  $X_k$ . In particular they may be Taylor expansions, in which  $B$ , the square matrix  $\|b_{k\alpha}\| = \|\partial X_k / \partial x_\alpha\|$ , and  $m$ , the column matrix  $\|m_k\| = \|\partial X_k / \partial \mu\|$ , are determined for the motion (5).

Note that in setting up the basic problem it is not necessary to assume that the motion (5) is a solution of some variational problem. The motions  $\eta$ ,  $\xi$  are constrained by the inequalities

$$R(x^* + \eta, \mu^* + \xi, t) \geq 0 \quad (8)$$

For any  $\eta(0)$ , perhaps constrained by some condition

$$\sum \eta_k^2(0) \leq A \quad (9)$$

where  $A$  is a given number, and for a given function  $\xi(0)$ , the relations (7) and (8) define what is called in automation a *transient process*.

One of the main problems of an automation specialist consists in the *control of the transient process*. He has to find the *technique and means for damping this process in what is in some sense the best manner*. In particular it is desirable that for  $t = T$

$$\sum \eta_k^2(T) \leq \delta \quad (10)$$

where  $\delta$  is a number chosen in a given way.

In order to achieve this, one may have at ones disposal a choice for the function  $\xi$ . Hence a certain measure is introduced for assessing the deviation of the actual motion (6) from the programmed one (5).

Let this be the functional

$$J(\xi) = r \int_0^T W(\eta, \xi, \dot{\xi}, t) dt + sw [\eta(T), \xi(T), T] \quad (11)$$

Here  $W$  and  $w$  are certain non-negative and sufficiently smooth functions of their arguments, while  $r$  and  $s$  are non-negative numbers.

The functional (11) is the *evaluation of the response of the transient process and of the final state of the system*. In order to complete the setting of the basic problem, one needs to choose the

\* Take also  $t_i = 0$ ,  $t_f = T$ .

class of comparison functions for which the functional (11) is determined. Mathematically it is convenient to choose this class as broadly as possible. The range of choice may, however, be limited by considerations of a practical engineering nature.

The simplest solution is formulated without any allowance for the form and characteristics of the modern technical equipment on which the achievement of the required solution will be based. Hence after formally defining such a class  $G$  of these functions (e.g.  $\xi, \eta$  to be smoothed-segment functions) the following basic problem on the search for an optimal regulator is stated.

**Fundamental Problem**—For the class  $G$  of permissible comparison functions it is required to determine a differential equation

$$F[\dots \dot{\xi}, \xi, r, t] = 0 \quad (12)$$

expressing the control law, such that a regulator constructed according to law (12) and connected to the plant (7) should ensure a minimum of functional (11) for all motions starting within (9).

### Note 1

The inequality (10) is not, generally speaking, a boundary condition. The expression  $\sum \eta_k^2(T)$  is itself a functional, and (10) merely expresses one's natural wish to bring the current point of the system into a certain limited domain of its state  $f$ . In the particular case where  $b_{k\alpha}, m_k = \text{const.}$  or are periodic functions in  $t$ , the ideal substitution  $T = \infty$  is permissible for functional (11).

Then one may put  $\delta = 0$ , and condition (10) will mean

$$\eta_k(\infty) = 0 \quad (k=1, \dots, n) \quad (13)$$

i.e. the regulator (12) must in addition ensure the asymptotic stability of the system. The latter is a concomitant of the finite nature of functional (11) for  $T = \infty$ .

If nevertheless  $T < \infty$ , then by putting

$$r = \frac{T}{T-t} - 1 \quad (14)$$

the basic problem of synthesis is brought into the form:

$$\begin{aligned} \dot{\eta} &= \frac{T}{(1+\tau)^2} (B\eta + m\xi + \Xi(\eta, \xi, t)) \\ J &= r \int_0^T \frac{T}{(1+\tau)^2} W(\eta, \xi, \dot{\xi}, \tau) d\tau + sw \end{aligned} \quad (15)$$

Hence for finite  $T$  also one may speak formally of the stability of the system as ensured by the regulator

$$F\left[\dots \frac{d\xi}{d\tau}, \xi, \eta, \tau\right] \quad (16)$$

This will essentially signify the monotonic damping of a certain positive definite function  $V(\eta, \xi, t)$  for  $t \in [0, T]$ .

In this case one can ensure that there exists a number  $\delta \leq A$  for which inequality (10) is satisfied. Of course such a transformation is otherwise of a formal nature, and does not essentially supersede the investigation of solutions of (15) and (16).

## Note 2

If  $W$  contains no derivatives of  $\xi$ , then (12) degenerates and becomes algebraic:

$$F[\xi, \eta, t] \quad (17)$$

One can say that this defines an *ideal regulator*, i.e. one which requires an energy source of infinite power.

Note that *formally*, by suitable substitution of the variables, the basic synthesis problem may sometimes be reduced to the achievement of a regulator of type (17) for all  $W$ . Given below is such an example.

## Note 3

In the statement of the fundamental problem it is assumed that the mathematical model (7) possesses the property of 'complete controllability and complete observability', or the less powerful property of 'complete stabilizability' (see under 'General Control Layout').

An example of the synthesis problem showing how the requirement of 'complete observability' may be reduced is now given. As for the requirement for 'complete controllability' or 'stabilizability', it appears to be so natural and self-evident that there cannot be any question of abating it.

In fact, the model (7) could surely only prove uncontrollable if some essential terms had been left out of account in eqns (1) or if the programme (5) had been chosen unreasonably.

Of course nobody can ensure against such an error, but this once more underlines the wisdom of referring the programming problem to the competence of a dynamics specialist. The 'controllability criterion' or the 'stabilizability criterion' presents him with an effective and rigorous means of finding out whether the model he has constructed is adequate for the given controlled plant. Only when the criterion is satisfied can he pass this model over to the automation specialist for further investigation.

## General Control Layout

Let it be assumed that the control law (12) is known. The general structural layout of the control system is shown in Figure 1.

Here the brachistochronic motion is contained in the pro-

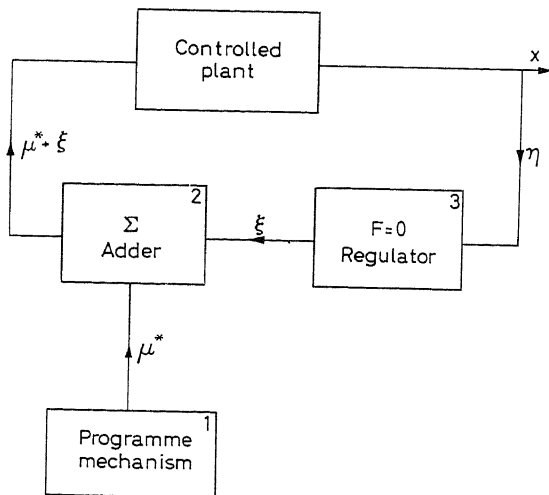


Figure 1. Overall control layout

gramme mechanism 1. Such a mechanism can always be constructed using a clock.

The signals from 1 pass to the summing device 2. They give rise to a programmed deviation of the controlling device  $\mu^*$ ; its additional deviation  $\xi$  is generated by the regulator 3, which uses information  $\eta(t)$  on the actual state of the controlled plant.

This information is detected by a suitable collection of measuring instruments operating according to the algorithm contained in (12). This layout expresses the general concept of modern automatic control theory.

## Possible Allowance for Various Characteristics of Modern Technical Equipment

Another statement of the problem of synthesis of optimal regulators relies on the following considerations.

For example, whatever the form of the optimizing functional (11) and the nature of eqn (12), one can only achieve this equation by the use of a servomotor, whose speed is limited and depends appreciably on the external load.

In such cases it may be better, for example, to take account at once of the equation

$$\dot{\xi} = f(\sigma) \psi(w) \quad (18)$$

for the speed of the servomotor, if it is hydraulic. Here  $\psi(w)$  is the prescribed function

$$\psi = \sqrt{w} \quad (19)$$

for reversible servomotors and

$$\psi(w) = \begin{cases} 1 & w \geq 1 \\ \sqrt{w} & 0 < w < 1 \\ 0 & w \leq 0 \end{cases} \quad (20)$$

for non-reversible servomotors.

In the general loading case

$$w = 1 - (P^2 \ddot{\xi} + Q \dot{\xi} + r \xi) \text{ sign } \sigma \quad (21)$$

Here  $P$ ,  $Q$  and  $r$  are given constants.

As for the function  $f(\sigma)$ , it is only defined to the extent that it belongs to a certain class  $B$  of functions characterized by the properties:

$$\begin{aligned} (1) \quad & f(\sigma) = 0, \quad \sigma f(\sigma) > 0 \quad \text{for } |\sigma| \leq \bar{\sigma} \\ (2) \quad & f(\sigma) \equiv 0 \quad \text{for } |\sigma| > \bar{\sigma} \\ (3) \quad & |f(\sigma)| \leq \bar{f} \end{aligned} \quad (22)$$

Here  $\bar{\sigma}$  and  $\bar{f}$  are given positive numbers.

The class  $\eta$ ,  $\xi$  of smoothed-segment functions that satisfy eqns (7) and (18), conditions (9) and limits (8) and (22), can be described as permissible.

Problem: For the class of permissible functions, complete the construction of eqn (18), i.e. determine accurately the function  $f(\sigma) \in B$  and the switching function

$$\sigma = \sigma(\eta_1, \dots, \eta_n, \xi, t) \quad (23)$$

that make the functional (11) a minimum for all motions starting at (9) and make the system (7), (18) stable.

The problem may be formulated similarly to take into account the characteristics of an electrical servomotor or of the measuring equipment used in the regulator.

### Selection of the Optimizing Functional

The selection of the optimizing functional is made on the basis of the view as to what is 'good' and what is 'bad'. Beyond doubt this question should be resolved in the light of experience. In a number of particular cases one considers the functionals

$$\int_0^T W(\eta, \xi, \dot{\xi}, t) dt \quad (24)$$

which assess the extent of smoothing of the perturbed motions  $\eta, \xi$  along brachistochrones. The presence in  $W$  of suitably chosen expressions in  $\xi$  and  $\dot{\xi}$  has the aim of minimizing the magnitude of (24) with minimum energy consumption by the regulator.

In other problems—terminal control problems—one studies the functionals

$$w[\eta(T), \xi(T), T] \quad (25)$$

which represent the magnitude of deviation of the system from its terminal condition ( $f$ ) at the instant  $t = T$ .

So far only quadratic functionals have been studied. In all instances no rigorous grounds are given for the choice of the functions  $W$  and  $w$ . This choice is made to a considerable extent on an intuitive basis, and constitutes the first experiment in the construction of a theory of the synthesis of optimal regulators. From the formal point of view, the functional, if it is previously chosen, may be considered as a postulate which determines the nature of the control theory deriving from it. Hence, just as happens in geometry, each functional assumed *a priori* will have, corresponding to it, its own theory of optimal regulator design. Just as from all the possible geometries at the outset of its development mankind chose for its first science of measurement the simplest and intuitively closest geometry of Euclid, so today the theory of the synthesis of optimal regulators is being constructed round the quadratic functional as the simplest and intuitively closest.

There are also other more prosaic yet important reasons for investigating quadratic functionals.

As will be seen, for example, the well-worked-out technique for solving problems of regulator synthesis as optimized by quadratic functionals leads one directly to the solution of the same problems in which the functionals (11) are appreciably more complex.

### Regulators which are Optimal for Response Speed

If one puts  $\delta = 0$ ,  $rW \equiv 1$ , and  $s = 0$  in (10) and (11), one arrives at the problem of the rapid-response regulator.

Though this problem fits into the general framework of the basic problem, it still occupies a somewhat special place which justifies its separate treatment in what follows.

### Progress Achieved in Solving the Basic Problem

#### General Situation

The synthesis problem is not as deeply rooted in science as the programming problem. It originates in the demands of modern engineering, and is caused by the historical course of its development.

Engineers, by relying on experience and intuition, have designed (as far as they were able) regulators first of type (17) and then of type (12), mainly in order to maintain required steady-state operating conditions in controlled plants.

Their intuition has led to the discovery of two basic principles of control: *those of control by displacement ( $r$ ) and control by perturbation ( $f$ )*. The discovery of the feedback principle and its wide application in practice has played an immense part in the development of automatic control systems.

Essentially, science is now occupied in seeking merely more accurate and more complete forms of these principles. The simplest example of such an improvement is the theory of observability which is now being developed; in more complex cases, in accordance with Wiener's cybernetic doctrine, one is endeavouring to find comprehensive forms of these principles in systems with adaptation.

In itself the wish to discover optimal forms is not so much praiseworthy as necessary, in view of the extremely limited material resources at the disposal of mankind.

One is bound to think of the future generation and to expend ones resources as economically as possible.

There are a number of instances, few but nevertheless impressive, in which an effective solution of the synthesis problem has been or can be obtained. These will now be described.

First, however, the characteristics of the methods employed will be briefly dwelt upon. Here reliance is placed on the evidence existing in the literature for showing what these methods have been able to yield.

Sincerest apologies must be offered to mathematicians for so primitive an approach to the assessment of the results of their creative activity. For those occupied in the application of mathematical methods such an approach is inevitable and is justified by the desire to obtain as soon as possible in analytic form an expression of the concrete content of the synthesis problem.

The assessment of the potentialities of these methods remains, however, the province of the mathematicians themselves. The most that one can indicate here consists merely of an assembly of perhaps inadequately justified guesses.

### Methods for Solution of the Synthesis Problem

The first and simplest synthesis problems were solved by the techniques of the calculus of variations. These techniques have been extensively developed in the works of the mathematical school founded by Pontryagin, and they have been given elegant expression in the form of the general theorem known as the maximum principle.

The maximum principle has made for improved understanding and solution of complex modern applied variational problems defined in closed spaces. Variational techniques also possess other important advantages that do not exist in other techniques. For example, they enable one to establish the following fact quite simply. If the function  $W$  in (11) contains a derivative of  $\xi$  of order  $m$ , then the eqn (12) will also be of order  $m$ . The practical application of these techniques however runs up against difficulties, which consist in the need to integrate the equations of the variational problem. The difficulties of integration are increased by the fact that  $n$  auxiliary Lagrange multipliers are normally employed in these techniques. The latter are completely unconnected with the essential content of the problem, and do not enter at all into its statement.

On this path the famous two-point boundary problem stands as a fortress that has been attacked time after time but never conquered. In those rare instances where it becomes possible to express the multipliers  $\lambda$  in terms of the variables  $\eta, \xi$  and  $t$ , the variational techniques work successfully.

The author, however, is not so bold as to claim that success is close at hand, in many other cases where the application of variational techniques appears both rigorous and legitimate. In a number of particular cases the use of the method of dynamic programming seems more attractive.

If one can make use of the definition of  $\delta$ -functions, then the method of dynamic programming leads one to the investigation of partial differential equations in the function

$$V(\eta(t), \xi(t), t) = \int_t^T [rW + s\sigma(t-T)\omega] dt \quad (26)$$

This equation serves as a mathematical expression of the Bellman optimization principle.

The optimization principle is extremely broad, and the method of dynamic programming which is based on it is applicable to the investigation of optimal processes of very broad scope. Of course this method introduces quite a number of difficulties of various sorts. For example, the application of the method to the solution of the basic problem is limited in extent to functions  $V$  which are continuous and smoothed-segment in  $\eta$  and  $\xi$ , and it is impossible to find out in advance whether one is dealing with these cases. Fortunately in its application to this particular branch of science the dynamic programming method is closely related to Liapunov's direct method of determining  $V$  functions to solve the stability problem. The link is as follows. As a rule it turns out that the optimization principle is satisfied by a certain set of optimizing functions  $V(\eta, \xi, t)$ . However, only those functions of this set that are Liapunov functions for the closed system will provide an actual solution of the problem.

Regulators working according to such functions are optimal and also give the system closed-loop stability.

It follows that the Liapunov method enables one to select the required solutions out of those that are provided by the Bellman optimization principle. This combination of the two techniques also makes possible an effective solution of the synthesis problem. It assumes a special importance also because in the actual statement of the synthesis problem there is a great need for Liapunov's rigorously-based and fundamental concept of perturbed and unperturbed motions.

Essentially this concept appeared spontaneously as a basis for the development of classical control theory, while now it is also becoming the foundation on which modern control theory is being built.

It is thought that the Bellman-Liapunov method has its widest possibilities in those cases where some form of successive approximation technique is applied to it. One can now discuss the results obtained by this method.

The fundamental equations of the method set out in the form of sufficient conditions for optimization are now given.

Let the equations of perturbed motion be:

$$\dot{\eta} = f(\eta, \xi, \tau) \quad (27)$$

These are defined for  $\|\eta\| < R$ ,  $\tau \in [0, \infty]$ ;  $R < \infty$  is a given number.

By choice of the differential eqn (12) in  $\xi$  it is required to minimize the functional

$$J = \int_0^\infty W(\eta, \xi, \tau) d\tau \quad (28)$$

such that the motion  $\eta \equiv 0$  shall be asymptotically stable relative to initial perturbations from the region  $\|\eta_0\| < R_0 \leq R$ . It may happen that  $R_0 = R = \infty$ . If there exists a function  $V(\eta, \tau)$  and a function  $\xi^0$  satisfying eqn (17), such that:

(1) they are defined in the region  $\|\eta\| < R$ ;

(2) the function  $V(\eta, \tau)$  is positive definite;

(3) at the point  $\eta \equiv 0$  the function  $V$  assumes an infinitely small upper bound.

(4)  $\sup [V(\eta, \tau) \mid \|\eta\| \leq R_0] < \inf [V(\eta, \tau) \mid \|\eta\| = R]$

(5) The derivative  $\dot{V}$  calculated along the trajectories of the system as closed by the regulator (17) satisfies

$$(\dot{V})_{\xi^0} = -W(\eta, \xi^0, \tau) \quad (29)$$

and the optimization equation

$$(\dot{V})_{\xi^0} + W(\eta, \xi^0, \tau) = \min_{\xi} [(\dot{V})_{\xi} + W(\eta, \xi, \tau)] \quad (30)$$

Then  $\xi^0$  is an optimal control, and

$$V(\eta_0, 0) = \min_{\xi} \int_0^\infty W(\eta, \xi, \tau) d\tau \quad (31)$$

is valid. If eqn (17) has been found, then the stability of the system is ensured by conditions (1)–(5), since the  $V$  thus determined is a Liapunov function, while the optimal nature of the regulator follows from eqn (30), which expresses the Bellman optimization principle.

For  $T < \infty$ , the principle as formulated retains its meaning provided that in defining the function  $V$  one brings in also the boundary condition

$$V(\eta, T) = sw \quad (32)$$

and treats the word 'stability' in the sense explained under the heading 'Synthesis of Optimal Regulators'.

It is not necessary to dwell on the characteristics of the functional analysis techniques and the various types of approximate methods that have been successfully applied to the solution of a number of variational problems.

## Synthesis of Ideal Regulators

### Stationary Linear Systems. Problem 1

*Statement of the Problem*—Consider the stationary linear systems (7) ( $\mathcal{E}_k \equiv 0$ ), defined for  $\mu^* \equiv 0$ ,  $T = \infty$  and

$$|\xi| \leq \bar{\xi} \quad (33)$$

where  $\bar{\xi}$  is a given number.

The functional (28) is calculated for

$$rW = W_2 = \sum \sum a_{kj} \eta_k \eta_j + c\xi^2 \quad (34)$$

Here the  $d_{kj}$  and  $c$  are positive weighting coefficients.

The optimal solution follows from eqns (30) and (31). The Liapunov function is a quadratic form

$$V_1 = \sum \sum A_{kj} \eta_k \eta_j \quad (35)$$

in which the  $A_{kj}$  are real solutions of the system of algebraic equations

$$a_{kj} + 2 \sum b_{ak} A_{aj} - \frac{1}{c} (\sum m_{ak} A_{ak}) (\sum m_{aj} A_{aj}) = 0 \quad (36)$$

satisfying Sylvester's criterion.

The equation of the regulator is

$$\bar{\zeta} = -\frac{1}{c} \sum m_k \frac{\partial V_1}{\partial \eta_k} = \begin{cases} +\bar{\zeta} & \text{for } \sigma \geq +\bar{\zeta} \\ \sigma & \text{for } |\sigma| < \bar{\zeta} \\ -\bar{\zeta} & \text{for } \sigma \leq -\bar{\zeta} \end{cases} \quad (37)$$

Here  $\sigma$  is the linear function

$$\sigma = -\frac{1}{c} \sum p_\alpha \eta_\alpha \quad (38)$$

It has coefficients

$$p_\alpha = \sum m_k (A_{k\alpha} + A_{\alpha k}) \quad (39)$$

which play the part of gain coefficients.

Problem 1 is unsolved for  $c = 0$ .

#### Optimization with a Non-stationary Functional. Problem 2

With the aim of getting a transient process in Problem 1 having more intensive damping of perturbed motions, one puts in (28)

$$rW = X(t) W_2 \quad (40)$$

Here  $X(t)$  is an increasing function. For practical purposes it is convenient to take

$$X(t) = e^{2\delta t}, \quad \delta > 0 \quad (41)$$

In this case the solution (37) retains its form. Only now the Liapunov function is

$$V_2 = e^{2\delta t} V_1 \quad (42)$$

while the coefficients  $A_{kj}$  are solutions of the Riccati equations

$$\begin{aligned} \dot{A}_{kj} + 2\delta A_{kj} + a_{kj} + 2 \sum b_{ak} A_{aj} - \\ - \frac{1}{c} (\sum m_\alpha A_{\alpha k}) (\sum m_\alpha A_{\alpha j}) = 0 \end{aligned} \quad (43)$$

The solutions are selected in accordance with Sylvester's criterion.

In case (41) one may put  $A_{kj} = \text{const.}$

The point about the more intensive damping of the transient process is as follows:

Since the characteristic index of  $X(t)$  is  $2\delta$ , and the integral  $\int_0^\infty X(t) W_2 dt$  must converge, thus the characteristic index of  $\eta_k$  must always be strictly greater than  $\delta$ . Unfortunately for a regulator designed in this manner the integral (28) becomes a monotonic increasing function of  $\delta$ . This expresses the fact that one's wish to obtain a transient process with more intense damping of the oscillations cannot be matched harmonically to the condition for minimizing the energy consumption of the regulator, as expressed by the term  $c$ .

Practically, satisfactory damping results may be achieved with sufficiently large  $\delta$ ; the increments in the value of the functional will be insignificant provided the constant  $c$  is small enough.

The above problem may be generalized within the framework of the same technique if the weighting coefficients are taken as given functions of time.

#### Non-stationary Linear Systems. Problem 3

One returns again to the linear systems (1), in which

$$b_{k\alpha} = b_{k\alpha}(t), \quad m_k = m_k(t), \quad t \in (0, \infty) \quad (44)$$

are known functions of time.

The functional (28) will be optimized in which  $rW$  is defined by (40).

In this form the solution (27) remains valid, but now the eqns (43) cannot be satisfied by  $A_{kj} = \text{const.}$ , and one is faced with the problem of seeking its solutions for  $t \in [0, \infty]$  that satisfy Sylvester's criterion. This problem is rather difficult and requires a special discussion.

#### Non-stationary Linear Systems Defined over a Finite Interval. Problem 4

If the  $b_{k\alpha}$  and  $m_k$  are defined for  $t \in [0, T]$ , Problem 4 may be reduced to Problem 3 by employing the transformation (14).

A more direct path will be followed.

Therefore consider the functional

$$J = \int_0^T e^{2\delta t} W_2 dt + S \sum \sum C_{kj}(T) \eta_k(T) \eta_k(T) \quad (45)$$

Here  $S$  and  $T$  are fixed positive numbers, and the coefficients in  $W_2$  are known functions of time; the  $C_{kj}$  satisfy Sylvester's criterion.

The optimal solution is provided by a Liapunov function of form (42). As before, the coefficients of  $V_2$  satisfy the eqns (43). For these equations one has to solve Cauchy's problem with  $t \in [0, T]$ . The initial values of the variables are given for  $t = T$ , and in the integration time  $t$  varies backwards.

In accordance with condition (43), the initial values are determined by the formulae

$$A_{kj}(T) = s_j C_{kj}(T) \quad (46)$$

It is of interest to note that if eqn (17) exists, then it is continuously dependent on the parameter  $T$ , and in the case where the  $b_{k\alpha}$ ,  $m_k$ ,  $d_{kj}$  and  $c$  are constant at the limit as  $T \rightarrow \infty$ , it tends to the equation appearing in Problem 1. This opens the way to the solution of Problem 1 on an electronic simulator. In fact, to obtain a solution of Problem 1 one must choose eqns (43) on an electronic simulator, inverting the time according to  $\tau = T - t$ , and solve them for sufficiently large  $T$  with zero initial conditions. The values  $A_{kj} = \text{const.}$  which provide solutions to Problem 1 are stable as  $\tau \rightarrow \infty$ . Hence all solutions of eqns (43) that are found for  $A_{kj}(T) = 0$  will tend to  $A_{kj}$  as  $\tau$  increases.

#### Synthesis of Regulators Supplied from an Energy Source of Limited Power

##### Regulators with Finite Actuator Velocity. Stationary Systems. Problem 5

To return to Problem 1. Here one condition only will be changed: the optimizing functional. In fact, instead of the function (34) consider the function

$$rW = W_2 + \alpha \dot{\zeta}^2 + C \dot{\zeta}^2, \quad \alpha > 0 \quad (47)$$

Application of the Bellman-Liapunov technique leads to the following equation for the regulator:

$$\begin{aligned} 0 & \quad \text{for } \frac{\sum q_\alpha \eta_\alpha}{r} \geq +\bar{\zeta} \\ \dot{\zeta}^0 = \sum q_\alpha \eta_\alpha - r\dot{\zeta} & \quad \text{for } \left| \frac{\sum q_\alpha \eta_\alpha}{r} \right| < \bar{\zeta} \\ 0 & \quad \text{for } \frac{\sum q_\alpha \eta_\alpha}{r} \leq -\bar{\zeta} \end{aligned} \quad (48)$$

The gain coefficients  $q_\alpha$  and  $r$  are found by the solution of equations similar to (43). One encounters them in Problem 6.

Recently A. I. Lur'ye has obtained interlocked formulae for computing  $p_\alpha$ ,  $q_\alpha$  and  $r$  in terms of the parameters  $b_{k\alpha}$ ,  $m_k$  and the weighting constants of the functional.

*Regulators with Limited Servomotor Velocity. Stationary Systems. Problem 6*

If in Problem 5 one requires definitely to make allowance for the fact that the servomotor velocity is limited and may only be chosen from the class  $B$  of functions  $f(\sigma)$ , while at the same time there is no limitation on  $\xi$ . One may exercise this choice in the stationary case, keeping to the statement of the problem propounded on p. 251.

Here one can not only determine the form of the servomotor characteristic  $f(\sigma) \in B$  precisely, but can also find the form of the switching function  $\sigma(\eta_1, \dots, \eta_n, \xi, t)$ .

For the case  $w \equiv 1$ ,  $\bar{\sigma} = \infty$ , and a quadratic functional of type (11) in which

$$rW = W_2 + c\xi^2 + \alpha\xi^2, \quad \alpha > 0 \quad (49)$$

the servomotor equation is as follows

$$\begin{aligned} +\bar{f} & \quad \text{for } \sigma \geq +\frac{\bar{f}}{h} \\ f(\sigma) = h\sigma & \quad \text{for } |\sigma| < \frac{\bar{f}}{h} \\ -\bar{f} & \quad \text{for } \sigma \leq -\frac{\bar{f}}{h} \end{aligned} \quad (50)$$

Here  $h$  is a given positive constant, while  $\sigma$  is defined by

$$\sigma = -\frac{1}{2\alpha h} (\sum B_k \eta_k + 2R\xi) \quad (51)$$

This solution is derived from the Liapunov function:

$$V_3 = \sum \sum A_{kj} \eta_k \eta_j + \xi \sum B_k \eta_k + R\xi^2 \quad (52)$$

The coefficients  $\tilde{V}_2$  are solutions of the equations

$$\begin{aligned} a_{kj} &= 2 \sum b_{ak} A_{aj} - \frac{1}{4\alpha} (\sum b_r) (\sum B_s) = 0 \\ \sum_\alpha b_{ak} B_\alpha + 2 \sum_\alpha m_\alpha A_{aj} - \frac{R^2}{\alpha} (B_k) &= 0 \\ R^2 &= \alpha (c + \sum m_k B_k) \end{aligned} \quad (53)$$

$V_3$  being positive definite.

Problems 5 and 6 may be solved by introducing the new variable  $\zeta$ , and from there by adding to the model (7) the equation  $\dot{\zeta} = \zeta$  one can synthesize the ideal regulator:

$$\zeta = \zeta(\eta_1, \dots, \eta_{n+1}t) \quad (54)$$

This approach may not be employed, however, if one wishes to take account in the servomotor characteristic of its dependence on the load parameters, in the way that this was done in formulae (18)–(23).

For  $\alpha = 0$ , Problem 6 is variable.

*Optimization with a Non-stationary Functional. Problem 7*

One can repeat the solution of Problem 6, changing in it only the form of the functional.

In fact one puts in (28)

$$rW = e^{2\delta t} (W_2 + c\xi^2 + \alpha\xi^{02}) \quad (55)$$

The equation of the regulator retains the form (50), as the Liapunov function now appears as

$$V_4 = e^{2\delta t} V_3 \quad (56)$$

Its coefficients satisfy the Riccati differential equations:

$$\begin{aligned} \dot{A}_{kj} + 2\delta A_{kj} + a_{kj} + 2 \sum b_{ak} A_{aj} - \frac{1}{4\alpha} (\sum B_r) (\sum B_s) &= 0 \\ B_k^0 + \sum_0 b_{ak} B_\alpha + 2 \sum m_\alpha A_{aj} - \frac{R^2}{\alpha} (B_k) &= 0 \\ \dot{R}^2 + R^2 &= \alpha (c + \sum m_k B_k) \end{aligned} \quad (57)$$

For stationary systems one can satisfy eqns (57) by putting  $A_{kj}$ ,  $B_k$ , and  $R = \text{const}$ . For  $\delta = 0$  the solutions coincide with those of eqns (53).

The solutions must be selected in accordance with Sylvester's criterion.

*General Case of Non-stationary Linear Systems. Problem 8*

Retaining all the assumptions of Problem 7, the optimizing functional only is changed. One puts

$$\begin{aligned} J &= \int_t^T e^{2\delta t} (W_2 + c\xi^2 + \alpha\xi^{02}) dt + s [\sum \sum C_{kj}(T) \eta_k(T) \eta_j(T) + \\ &+ \xi(T) \sum \alpha_j(T) \eta_j(T) + e(T) \xi^2(T)] \end{aligned} \quad (58)$$

Here  $S$  and  $T$  are fixed positive numbers, and  $C_{kj}$ ,  $d_j$  and  $e$  are given numbers for which the corresponding form is positive definite. The optimal solution is provided by a Liapunov function of form (56) whose coefficients satisfy the eqns (57). For these equations one has to solve CAUCHY's problem over the interval  $[0, T]$ , taking as initial conditions

$$\begin{aligned} A_{kj}(T) &= sC_{kj}(T) \\ B_k(T) &= sd_k(T) \\ R(T) &= se(T) \end{aligned} \quad (59)$$

Equations (57) may be used for solving the corresponding stationary problem, Problem 6 by the method described in Problem 4.

*Isoperimetric Conditions. Problem 9*

The basic problem may be complicated by isoperimetric conditions which represent some additional requirement placed on the system.

For example, the limitation

$$m \int_0^T \xi^2 dt + n \int_0^T \xi^{02} dt \leq l, \quad m, n \geq 0, \quad l > 0 \quad (60)$$

represents the desire to achieve the best possible smoothing of the transient process under conditions where the actuator power consumption is strictly limited and so is its available displacement. In (34) and (40) one may put  $c, \alpha \geq 0$ .

In other instances the isoperimetric conditions may concern limitations in characteristics of a more general nature, including the coordinates of the controlled plant.

In particular, for functional (11) with  $T$  previously specified, condition (10) may be considered as isoperimetric. Isoperimetric synthesis problems have scarcely been studied. Only for the simplest stationary case treated in Problem 1 has the following been established.

Let it be assumed that in Problem 1 one has  $c=0$  in the function  $W_2$  (34), while the limitation (60) with  $n=0$  is applied to the motion of the actuator. Then it has been shown that the solution of the isoperimetric problem does not differ from solution (35) of Problem 1. Its special feature consists in the fact that the isoperimetric condition (60) leads to the appearance in the phase space of a closed region of initial values  $\eta_0$ , containing the origin, for which the solution (34) exists.

It has been shown that provided the initial perturbations are sufficiently small, the isoperimetric limitation is more powerful than limitation (33), i.e. for such perturbations the boundary in (33) is not reached for any  $t > 0$ .

#### Allowance for Perturbation Forces. Problem 10

It may happen that the programme (5) will be determined approximately. This will be the consequence of either of the following reasons:

(1) The solution (5) has been determined by numerical integration methods.

(2) The eqns (1) take no account of the forces that are present in practice.

In both cases eqns (7) will contain additional terms  $f_k(t)$ .

$$\dot{\eta} = B\eta + m\dot{\xi} + f(t) \quad (61)$$

where  $f(f_1, \dots, f_n)$  is a limited and vanishing vector function.

In many cases these are known limited time-functions, vanishing at infinity, which represent determinate perturbing forces in (47).

The presence of such functions gives one a basis for a statement of the synthesis problem for optimal regulators which takes into account the known information on the forces  $f(t)$  and improves on eqn (12) as previously found without making allowance for these forces.

The Liapunov function has the form

$$V = \sum A_{kj} \eta_k \eta_j + \sum B_k(t) \eta_k + D(t) \quad (62)$$

Here the first term coincides with  $V_1$ ; the values of  $B_k(t)$  and  $D(t)$  are determined from the known  $f_k(t)$ , and they are vanishing functions provided the  $f_k(t)$  are vanishing.

The sign of  $V$  is determined by its first term. The equation of the regulator splits up into two parts: in the first it coincides with eqn (37); while in the second it contains terms dependent on the  $f_k(t)$ . The regulator operates on the displacement principle and the perturbation principle. It achieves a lower value of the optimizing functional than would be obtained if information on the perturbation forces were not made use of. The very simple problem considered is capable of effective generalization to more complex cases.

#### Effect of Slight Non-linearities. Problem 11

In a number of applied problems it is extremely desirable to allow for the effect produced by non-linear terms  $\Xi(\eta, \xi, t)$ .

Now study the treatment of the simplest case. One assumes that eqns (7) are stationary, that the non-linear terms are in-

dependent of  $\xi$ , and that the following expansion in the coordinate  $\eta_k$  exists in integral convergent series:

$$\dot{\eta}_k = \sum_{\alpha} b_{k\alpha} \eta_{\alpha} + m_k \dot{\xi} + \sum_{\alpha} \sum_{\beta} C_{\alpha\beta}^{(k)} \eta_{\alpha} \eta_{\beta} + \dots \quad (63)$$

Returning to Problem 1, and retaining all the facets of the problem except that instead of initial linear equations one now considers the eqns (63).

An optimal  $V$ -function may be sought as a sum of the forms

$$V = V_2 + V_3 + V_4 + \dots \quad (64)$$

expanded in integral values of the variables  $\eta_k$ . For example,

$$V_m = \sum A_{\alpha, \beta, \dots, \omega} \eta_{\alpha}^{\alpha} \eta_{\beta}^{\beta} \dots \eta_{\omega}^{\omega} \quad (65)$$

where the sum is calculated over all integral indices  $\alpha, \dots, \omega = 1, \dots, n$  with the condition

$$\alpha + \beta + \dots + \omega = m \quad (66)$$

In attempting to satisfy eqns (30) and (31) with this form, one arrives at systems of recurrence equations for the coefficients  $A_{\alpha, \beta, \dots, \omega}^m$ .

Thus one obtains eqns (36) for the form  $V_2$ , which agrees precisely with form (35). It is a characteristic feature that if eqns (36) are soluble, then the systems of recurrence equations for determining the coefficients of the other forms are also soluble.

As before, the equation for the optimal regulator is determined from the formula

$$\xi = \frac{+\bar{\xi}}{-\bar{\xi}} \frac{1}{c} \sum_{m_k} \frac{\partial V}{\partial \eta_k} = \sigma, \quad \begin{matrix} \sigma \geq +\bar{\xi} \\ |\sigma| < \bar{\xi} \\ \sigma \leq -\bar{\xi} \end{matrix} \quad (67)$$

whose right-hand side now consists of an infinite series of integral powers of the coordinates  $\eta_k$ . The convergence of the series (64) and (67) is ensured, at least for sufficiently small  $||\eta||$ .

No difficulty arises in principle for applying the technique also to those cases where the right-hand sides of eqns (7) can be represented as integral series of a small parameter  $\varepsilon$ . The same rule holds good in all cases: the non-linear problem is soluble provided the linear corresponding problem is also soluble. Here lies the great value of investigating the synthesis problem for linear systems optimized by a quadratic functional.

#### Effect of Slight Non-linearities. Non-stationary Systems.

##### Problem 12

The assumption of the stationary nature of system (63) can be omitted from Problem 11. Eqns (30) and (31) may be satisfied by the function  $V$  (64) if one assumes that its coefficients  $A_{\alpha, \beta, \dots, \omega}^m$  are dependent on  $t$ . There will be systems of recurrence Riccati equations to determine them.

As before, the equation of the regulator is determined from the formula (67): but now the coefficients of the series will depend on  $t$ . As for the properties of the Riccati equations, one may draw a number of conclusions in this case which are similar to those derived in Problem 4.

##### More Complex Criteria Differing from the Quadratic. Problem 13

It must unfortunately be stated that no instance is known of the effective solution of synthesis problems in which the optimiz-

ing criterion is other than quadratic. Nevertheless, one problem is indicated here that is entirely convenient for systematic study by the techniques described above. The case in mind is that of an optimization criterion chosen such that the functions  $W$  and  $\omega$  may be represented by integral series of  $\eta$ ,  $\xi$  and  $\xi^0$ . The coefficients of the series may be time-dependent. In these cases one can develop the method of solution set out in problems 11 and 12, even if the controlled plant is non-stationary and contains non-linear terms  $\Xi(\eta, \xi, t)$ .

#### Regulator with Delay. Problem 14

In the following substantial development of the ideas of pp. 251—252 and Problem 6 for taking account of possible characteristics of the technical equipment employed, using the concept of incomplete observability, a most effective result has recently been obtained on the basis of the following assumptions:

(1) In the control process it is only possible to measure and store, during a limited time, the quantities  $W_1, \dots, W_l$  ( $l \leq n$ ), which are related to  $\eta$  by the vector equation

$$W = \phi(t, \eta), \phi\{\phi_1, \dots, \phi_e\} \quad (68)$$

which is not uniquely soluble for  $\eta$ .

(2) The measurement of the components of the vector  $W$  occurs with a time delay  $\theta$ .

(3) The time  $T = \infty$ .

Then the equation of the servomotor takes the form:

$$\begin{aligned} \xi^0 &= f[\eta(t+\theta), \xi(t+\theta), t] \\ -\tau \leq \theta \leq 0; 0 < \tau = \text{const.} \end{aligned} \quad (69)$$

In these assumptions one is seeking the precise form of the functions ' $f$ ' appearing in eqn (69), such that:

(1) The unperturbed motion of the closed system (7), (69) should be asymptotically stable.

(2) For all perturbed motions arising from (9) and  $\xi(0)$  the functional (11) should be a minimum, the  $W$  in it being a positive definite analytic function of the variables  $\eta$ ,  $\xi$  and  $\xi^0$ .

In particular, in the linear case, the required control law has the form:

$$\dot{\xi} = r(t)\xi + \int_{t-\tau}^t [N(t, \theta)W + M(t, \theta)\xi(t+\theta)]d\theta \quad (70)$$

Here the functions  $r$ ,  $N$  and  $M$  are continuous in  $t$  and  $\theta$  and uniformly bounded with respect to  $t$ .

A computing algorithm has been developed for determining them, which, in particular, includes a solution of the Riccati equations.

#### More General Case of Systems with Delay. Problem 15

Let equations of perturbed motion be considered under the form

$$\dot{\eta}_k^0 = f_k[\eta(t), \eta(t-\tau), \xi, t] \quad (71)$$

defined in a certain region  $N(\eta, \xi) > 0$  and  $t \in [0, T]$ .

One assumes that the motions are determined for a certain initial set of functions  $M$

$$\begin{aligned} \eta(t) &= g(t) \\ -\tau \leq t \leq 0 \end{aligned} \quad (72)$$

Consider now the functional

$$J = \int_0^T W[\eta(t), \eta(t-\tau), \xi(t), t] dt + Q[\eta(T), \eta(T-\tau)] \quad (73)$$

It is defined for the class  $G$  of continuous-segment  $\xi$  and smoothed-segment functions  $\eta$ . Obviously the control must depend on the state of the system at the instant  $t$  and also on its state over the preceding time interval  $[t+\theta, t]$ ,  $-\tau \leq \theta \leq 0$ . Hence the equation of the optimal regulator, if any, will be of the form:

$$\begin{aligned} F[\xi(t), \eta(t), \eta(t+\theta), t] &= 0 \\ -\tau \leq \theta \leq 0 \end{aligned} \quad (74)$$

The following problem arises. For the class  $G$  of permissible comparison functions, it is required to determine a regulator eqn (74) such that:

(1) it minimizes the functional (73) for all motions of the system starting from the set  $M$  (72).

(2) it makes the closed system (71), (74) stable.

The equations of pp. 250—252 are capable of generalization to systems of type (71), (74), and the regular procedure of applying them may be developed for all the particular cases of the problems described above.

Thus, for example, in the case of minimizing the quadratic functional  $W_2$  (34) in its application to a linear stationary controlled plant

$$\begin{aligned} \eta_k^0(t) &= \sum_{\alpha} (b_{k\alpha}\eta_{\alpha} + c_{k\alpha}\eta_{\alpha}(t+\theta)) + m_k \xi(t) \\ -\tau \leq \theta \leq 0 \end{aligned} \quad (75)$$

the problem is solved by a Liapunov function of the form

$$\begin{aligned} V &= \sum_k \sum_i \left[ m_{ki}\eta_k(t)\eta_i(t) + \eta_k(t) \int_{-\tau}^0 \beta_{ki}(t, \theta)\eta_i(t+\theta)d\theta \right] \\ &+ \int_{-\tau}^0 \int_{-\tau}^0 \left[ \sum_k \sum_i \gamma_{ki}(t, \theta, h)\eta_k(t+\theta)\eta_k(t+h)d\theta \right] dh \end{aligned} \quad (76)$$

The equation for the optimal regulator is:

$$\xi = -\frac{1}{c} \sum_k p_k \eta_k(t) + \sum_k \int_{-\tau}^0 (t, \theta)\eta_k(t+\theta)d\theta \quad (77)$$

Here the  $p_k$  are constants or functions of  $t$ , while the  $q_k$  are functions of  $t$  and  $\theta$ . Computation techniques have been developed for determining them.

It is appropriate here to mention an approximate method for solving the problem due to M. E. Salukvadze, based on an application of the Lagrange finite-increment formulae. The essence of the method is as follows. The delay interval  $[t-h, t]$  is split up into  $m$  equal sections, over which the function is described by means of the Lagrange formulae.

The formal introduction of new variables allows one to put this representation into the form of a system of linear differential equations. Then the original problem is completely replaced by another problem in which delay plays no part. Its solution is found by the methods described earlier. The value of the solution lies in its rapid approach to the accurate solution, as for example (76), (77), as the number  $m$  increases.



*Terminal Control. Problem 16*

Problem 16 follows from the basic problem pp. 247–248, if  $r = 0$  is put in (11). The system is optimized according to its terminal state at the instant  $t = T$ . Although this problem fits within the general framework of the methods described here, it has its own specific features and has been little developed.

The formulation of the problem is as follows (linear case).

Consider the perturbed motions described by the linear eqns (7). Here  $B(t)$  and  $m(t)$  are continuous matrices, defined for  $t \in [0, T]$ , while the  $\xi$  are continuous-segment functions satisfying the limitation (33).

Out of the class of permissible functions it is required to determine an eqn (17) for the regulator, for which the functional (11) attains a minimum for all motions starting at (9).

The problem calls for an application of Pontryagin's maximum principle method, which allows one to write down all the necessary relations for a solution. The control law is:

$$\xi = \bar{\xi} \operatorname{sign}(m^* \psi) \quad (78)$$

where the vector  $\psi$  satisfies the equation

$$\dot{\psi} = -B^* \psi \quad (79)$$

The asterisk denotes transposition of matrices. Success in applying the maximum principle is ensured by the fact that one can write down the following boundary condition for the vector  $\psi$ :

$$\psi(T) = -\operatorname{grad} w(\eta(T), T) \quad (80)$$

Consider now the space  $A$  of initial  $\eta_0$ , and the space  $B$  of terminal states  $\eta(T)$  of the controlled plant.

The equations for the problem enable one to set up a one-to-one correspondence between  $A$  and  $B$ . They also allow one to determine the instants  $t_1, \dots, t_m$  for the switchings of the actuator, the number of these, and the points  $\eta(t_i)$  corresponding to them in the space of states  $\eta$ .

Effective calculation of these relations can only be carried out using a computer, even in the simplest stationary cases.

*Regulators Optimized for Response Speed. Problem 17*

The following is a statement of the simplest problem in the synthesis of regulators optimized for response speed.

Consider a stationary linear system (7), defined for  $t \in [0, \infty]$ . Let  $0 < T < \infty$ , and determine the point

$$\eta(T) = 0 \quad (81)$$

If there exists a regulator (17) such that:

(1) it satisfies condition (33)  $|\xi| \leq \bar{\xi}$ ;

(2) when connected to the plant (7) it damps out any motion starting at (9), i.e. the solution

$$\eta(\infty) = 0 \quad (82)$$

is asymptotically stable;

(3) it brings the system to the position (81) in a minimum time  $T$ ;

then one can call such a regulator optimized for response speed.

The problem consists in determining eqn (17) of such a regulator from a class of suitably defined functions  $G$ . It falls within the outlines of the basic problem if in (11) one puts  $W \equiv 1$  and  $\omega \equiv 0$ . Its special feature is as follows. Since it is required in observing condition (33) to complete the transient

process in a finite time, one may assume in advance that the class  $G$  of permissible functions consists of continuous-segment  $\xi$  and smoothed-segment  $\eta$ . The eqns (30) and (31) for the problem are

$$-\frac{\partial V}{\partial t} = 1 + \sum_k \left( \sum_\alpha b_{k\alpha} \eta_\alpha + m_k \xi \right) \frac{\partial V}{\partial \eta_k} \quad (83)$$

$$\xi = \bar{\xi} \operatorname{sign} \sum_k m_k \frac{\partial V}{\partial \eta_k} \quad (84)$$

It would seem that if one were to solve eqns (83) and (84) and find the Liapunov function  $V$ , then the equation

$$\sum_k m_k \frac{\partial V}{\partial \eta_k} = 0 \quad (85)$$

would define the law for switching  $\xi$ . Here lies a difficulty, however, which one does not know how to overcome. In fact the representation (84) contains two equations, which must be satisfied by two functions  $V_1$  and  $V_2$ : that is, the equation

$$-\frac{\partial V_1}{\partial t} = 1 + \sum_k \sum_\alpha b_{k\alpha} \eta_\alpha \frac{\partial V}{\partial \eta_k} + \bar{\xi} \sum_k m_k \frac{\partial V}{\partial \eta_k} \quad (86)$$

$$\sum_k m_k \frac{\partial V}{\partial \eta_k} \geq 0 \quad (87)$$

and also the equation

$$-\frac{\partial V_2}{\partial t} = 1 + \sum_k \sum_\alpha b_{k\alpha} \eta_\alpha \frac{\partial V}{\partial \eta_k} - \bar{\xi} \sum_k m_k \frac{\partial V}{\partial \eta_k} \quad (88)$$

$$\sum_k m_k \frac{\partial V}{\partial \eta_k} \leq 0 \quad (89)$$

Obviously the functions  $V_1$  and  $V_2$  must assume identical values (by virtue of the continuity of  $V$ ) in

$$\sum_k \left( \frac{\partial V_1}{\partial \eta_k} - \frac{\partial V_2}{\partial \eta_k} \right) = 0 \quad (90)$$

A solution of the problem has not yet been achieved along these lines, and it is not known how to do it. To the engineer who is able to find a known answer in an approximate equation for a regulator optimized for response speed, one can only recommend the expedient of replacing the original precise problem by an approximate one.

One such expedient has been indicated in Problem 2; another consists in optimizing the functional

$$J = \int_0^\infty [(1-\theta)(W_2 + c\xi^2) + \theta\{1 - e^{-\lambda \sum_k \eta_k^2 + \xi^2 s}\}] dt \quad (91)$$

Here  $0 \leq \theta \leq 1$ ; for  $\theta = 0$  one gets Problem 1, while for  $\theta = 1$  one has an approximate version of the response-speed problem. The larger  $\lambda$  and  $s$ , the better is this version.

*Random Processes*

The range of synthesis problems broadens considerably if the model (7) [(61)] contains elements that vary randomly.

Such elements may be:

- (1) the programmed motion,
- (2) the parameters of the model,
- (3) the acting perturbation forces  $f(t)$ .

In another case one may wish to make allowance for the technical imperfection of the control equipment and to include the noise it generates in the synthesis problem.

If the probability characteristics of such random processes are known, the synthesis problem is formulated in terms of the theory of random functions. Essentially this problem has been conceived in the works of Kolmogorov, Khinchin and Wiener. They have laid the foundations of the modern theory of filtering. This theory is a large and independent section of the theory of automatic control, and has an extensive literature. The problem of synthesis of optimal systems with random properties may also be treated from the standpoint adopted in the present paper.

In fact, in a number of cases the use of the terminology of random function theory enables one to reduce the synthesis problem to a normal variational problem, into which one transfers the Bellman-Liapunov method of solution described here. This is demonstrated in the following example.

Let the equations

$$\dot{\eta} = f(t, \eta, \xi, x) + \gamma(t) \quad (92)$$

describe a random process (perturbed motion) starting at (9). Here  $\eta$ ,  $\xi$  and  $t$  have their previous meaning;  $x(t)$  is a limited Markov process representing the variation of a certain parameter, and  $\gamma$  is a noise. Their distribution functions are given. For the sake of simplicity we shall consider that  $\gamma \rightarrow 0$  if  $x \rightarrow 0$ . We assume that the equations are defined for  $\|\eta\| < R$ ,  $t \in [t_0, \infty]$ ,  $R$  being a given number. We introduce a definition of stochastic stability.

The motion  $\eta \equiv 0$  is called probabilistically stable if for any two numbers  $\varepsilon > 0$ ,  $p > 0$  there exists a  $\delta > 0$  such that the inequality

$$P\{\|\eta(t)\| < \varepsilon \text{ is satisfied for all } t > 0 / \|\eta(0)\| \leq \delta\} \geq 1 - p \quad (93)$$

Here  $P(\varphi|\psi)$  is the conditional probability of the event  $\varphi$  under the condition  $\psi$ . Furthermore, if the condition

$$\lim_{t \rightarrow \infty} P\{\|\eta(t)\| < \omega\} = 1$$

is satisfied for any  $\omega > 0$ , then the stability is said to be asymptotic.

Let a positive function  $W(\eta, \xi, t)$ , definite in  $\eta$  and  $\xi$ , and a positive definite function  $r(t, \eta)$  be given. Let the region of values of  $\eta$  that are permissible in the control process be defined by the inequality

$$\|\eta\| < R \quad (93a)$$

for all initial values from

$$\|\eta(0)\| \leq R_0, R_0 \leq R \quad (93b)$$

and we have to ensure that condition (93a) is satisfied with a given probability  $1 - q$ . Here  $0 < q < 1$  (a small  $q$  is desirable).

We shall characterize the response of the transient process by the conditional mathematical expectation

$$J(\eta_0, x_0, t_0) = M \left\{ \left[ \int_0^T W(\eta, \xi, \tau) d\tau + r(T, \eta(T)) \right] \middle| \eta_0, x_0 \right\} \quad (94)$$

$$\|\eta_0\| < R_0 \leq R$$

Here  $T$  is the random instant at which the achievement of the trajectory first attains the surface  $\|\eta\| = R_0$ .

*Problem:* for permissible functions find an equation

$$\xi = \xi^0(\eta, t, x)$$

of a regulator such that for  $\xi = \xi^0$ :

- (1) the functional  $J$  is bounded;
- (2) a minimum w.r.t.  $\xi$  is attained for any  $\eta_0$  from  $\|\eta(0)\| < R_0$  and any  $x_0$ ;
- (3) The solution  $\eta \equiv 0$  is probabilistically stable;
- (4)  $P\{\|\eta\| < R \text{ for all } t \geq t_0 / \|\eta(0)\| \leq R_0\} > 1 - q$  for any  $\eta(0)$  from  $\|\eta(0)\| < R_0$  and any  $x(t)$ .

To solve the problem, we employ the Bellman-Liapunov method in the following version.

If there exists in the region  $\|\eta\| \leq R$  a function  $v(\eta, x, t)$  and a function  $\xi^0(\eta, x, t)$  such that:

- (1)  $v$  is defined in  $\|\eta\| \leq R$ ;
- (2)  $v$  is positive definite;
- (3) at the point  $\eta \equiv 0$  there exists an infinitesimally small upper bound;
- (4)  $\sup [V_{\eta} v(\eta, x, t) | \|\eta\| \leq R_0] < \inf [V_{\eta} v(\eta, x, t) | \|\eta\| = R] q$
- (5) the averaged derivative  $dM\{v\}/dt$ , evaluated along the trajectories of the closed system, satisfies the equation

$$\frac{dM\{v\}}{dt} = -W$$

and the optimization condition

$$\left. \frac{dM\{v\}}{dt} \right|_{\xi^0} + W_{\xi^0} = \min_{\xi} \left[ \left. \frac{dM\{v\}}{dt} \right|_{\xi} + W \right] = 0$$

- (6)  $v(t, \eta, x) = r(t, \eta, x)$  on the surface  $\|\eta\| = R$ : then  $\xi^0$  is an optimal control, and

$$v = \min_{\xi} J$$

The function  $V(\eta, x, t)$  which satisfies these conditions is called an optimal Liapunov function.

There also exists a variant of the Bellman-Liapunov method for the case in which equations (2.6) are of more general form and are defined for  $t \in [t_0, T]$ .

This method enables an effective solution to be obtained in all the cases of given initial equations, optimizing functionals and characteristics of random processes that have been described above. Many of these results have already been derived. Both deterministic and stochastic processes admit optimization by the method described in many cases where they are of a discrete nature.

#### Processes with Adaptation

Methods have been outlined in the literature for the study of processes with adaptation. They are partly based on the Bellman optimization principle and the methods of probability theory. The question whether it is possible to develop a technique for studying processes with adaptation similar to that employed here cannot be answered now. One half of the method exists.

The immediate future will show whether an adequate second half of the method—some variant of Liapunov (or perhaps non-Liapunov) stability—will be evolved.

## Mathematical Aspects of the Synthesis Problem

### Existence of Solutions

The account in pp. 250-254 probably exhausts all the known particular synthesis problems that have been solved or can be effectively solved.

In this problem of the existence of the solutions themselves has not been considered. This extremely important, complex and interesting problem has been substantially developed in the theory of the maximum principle, in a number of papers by Bellman, Filippov, Kirillov, Kulikowski, and other authors.

For this purpose one could limit oneself entirely to an exposition of Kalman's highly elegant theory of the controllability of linear systems. To a considerable extent this theory is devoted to the solution of the existence problem. From the point of view adopted, however, in this paper, Krasovskiy's theory of stabilizability is more acceptable. In all publications the existence problem is reduced to the solution of two questions.

It is assumed that a permissible control exists. Then does the existence of an optimal control follow from this? The answer appears to be affirmative for a broad class of cases. Given this condition, the second question is: do permissible controls exist, and if so in what region  $N(x, \mu) > 0$  of the phase-space? The theory of stabilizability studies the eqns (7) in the neighbourhood of the origin. Let it be assumed that the equations do not depend explicitly on  $t$  and that the optimizing functional is quadratic.

Consider the vector subspaces corresponding to the eigenvalues  $\lambda$  of the matrix  $B = \text{const.}$ :

(1) The subspace  $S^+$  corresponding to the eigenvalues  $\lambda$  for which  $\text{Re } \lambda > 0$ ;

(2) The subspace  $S^0$  corresponding to the eigenvalues  $\lambda$  for which  $\text{Re } \lambda = 0$ ;

(3) The subspace  $S^-$  corresponding to the eigenvalues  $\lambda$  for which  $\text{Re } \lambda < 0$ ;

and the space  $R$  consisting of the vectors

$$m, Bm, \dots, B^{n-1}m$$

Then the following general result is obtained:

A. The system (7) is stabilizable no matter what the form of the non-linear terms  $\Xi(\eta, t)$ , provided the enclosure condition  $S^+ + S^0 \in R$  is satisfied. The law of the control  $\xi$  is a linear function of the type  $\xi = \sum p_j^* \eta_j$ .

B. The system (7) is not stabilizable by means of a linear control law no matter what the form of the non-linear terms  $\Xi(\eta, t)$ , if the enclosure condition is not satisfied and there exists at least one vector  $S_j^* \in S^+$  that does not belong to  $R$ .

C. One is obviously dealing with a critical case if the enclosure condition  $S^+ \in R$  is fulfilled but there is at least one vector  $S_j^* \in S^0$  that does not belong to  $R$ . In every critical case the question of the stabilizability of system (7) by the control  $\xi = \sum p_j^* \eta_j$  must be solved with reference to the nature of the non-linear forces.

This theory is capable of generalization to non-stationary systems, systems with delay, and non-quadratic functionals.

### Continuity of the Liapunov Function

The use of the Bellman-Liapunov method requires an examination of the fact that the function  $V$  is continuous in every problem. One returns now to Problem 1.

Here the region  $N(x, \mu) \geq 0$  is divided into three parts. For example, when  $|\xi| < \bar{\xi}$  the function  $V$  is defined by formula (35). Since  $V$  is positive definite, it represents a paraboloid in the  $\eta, V$  space. When  $\xi = \pm \bar{\xi}$ , the paraboloid is intersected by the planes

$$\frac{1}{c} \sum m_k \frac{\partial V}{\partial \eta_k} = \pm \bar{\xi} \quad (95)$$

which are parallel to the  $V$  axis.

The geometrical locus of the intersection points is described by a  $(n-1)$ -dimensional form.

Let it be assumed that  $p_n \neq 0$  (37).

From (95) is found

$$\eta_n = -\frac{1}{p_n} \left[ c(\pm \bar{\xi}) - \sum_{\alpha=1}^{n-1} p_\alpha \eta_\alpha \right] \quad (96)$$

This form is of the type:

$$V_2 = F[\eta_1, \dots, \eta_{n-1}, (\pm \bar{\xi})] \quad (97)$$

The new variable is introduced

$$\sum_{\alpha=1}^n p_\alpha \eta_\alpha + c(\pm \bar{\xi}) = \zeta \quad (98)$$

and reverts once more to the optimization equation.

If the system is in such a position that

$$\left| -\frac{1}{c} \sum m_k \frac{\partial V}{\partial \eta_k} \right| \geq \bar{\xi},$$

the actuator is on the stops  $\pm \bar{\xi}$ , and the equation for determining  $V$  becomes linear in the partial derivatives.

The coordinate  $\eta_n$  is eliminated and another linear equation is obtained in the partial derivatives:

$$\sum_{k=1}^{n-1} X_k \frac{\partial V}{\partial \eta_k} + X \frac{\partial V}{\partial \zeta} = f(\eta_1, \dots, \eta_{n-1}, \zeta) \quad (99)$$

where  $X_k$  and  $X$  depend only on  $\eta_1, \dots, \eta_{n-1}$ , and  $\zeta$ .

Thus one has arrived at the following Cauchy problem: it is required to find a solution  $V(\eta_1, \dots, \eta_{n-1}, \zeta)$  of eqn (99) which for  $\zeta = 0$  assumes the value  $V_2$  determined by formula (97).

Such a solution exists, at least in a certain neighbourhood containing the origin.

In the general case this problem remains open to investigation. Where there are discontinuities in the function  $V$ , the synthesis problem requires another statement taking into account the steps  $\Delta V$  on the lines of discontinuity.

### Conclusion

The author has tried to show that:

(1) The problem of synthesis of optimal regulators is fundamental to modern automatic control theory.

(2) Its rigorous statement is based on Liapunov's concept of perturbed and unperturbed motions.

(3) The physical and mathematical content of the problem is determined by the nature of the controlled plant and by the form of the optimization criterion.

(4) In many cases of deterministic and stochastic systems which are of practical interest in their own right, an effective solution of the synthesis problem has been and may be achieved by using both the methods of the calculus of variations and the Bellman-Liapunov method.

(5) In these cases it is of fundamental importance to solve the simplest problem on the optimization of a linear system by means of a quadratic functional.

(6) Effective criteria are now available for telling whether a synthesis problem is soluble or not.

(7) A persistent search can be seen in the literature for new methods of solution—functional analysis methods and approximate methods—and also for new cases where the problem is soluble (in the field of continuous, discrete, and stochastic systems), so that a clear outline is beginning to emerge of a well-built theory based on a unified method.

(8) The subject matter of investigations into the synthesis problem has good prospects of development in the next few years.

### Critique of Modern Automatic Control Theory

#### *A Particular Question in the Statement of the Optimal Control Problem*

It appears that by far the greater part of publications on optimal control theory is devoted to the problem of choosing brachistochronic motions (5). This is natural. It means that the boundaries of the region  $N(x, \mu) \geq 0$  are chosen to be extremely wide. This choice is not justified, and in a number of cases may make complete nonsense of the optimal control problem. Attention is drawn here to the existence of a deep connection between the problem of choosing a control programme and the basic problem. This connection exists independently of whether the programme (5) is a solution of some variational problem or whether it is any other manifestation of its creator's will. The connection is two-sided.

First, it is stressed once more that since the model (7) is defined for the motions (5), there is no *a priori* reason why it should be completely controllable or stabilizable.

Secondly, in all cases where the initial equations (1) are defined in a closed region  $N(\dot{x}, \mu) \geq 0$  the programmed motion (5) must not cross the boundaries of this region.

By using the simplest example the terminal control problem is now illustrated. It is assumed that the closed region is a layer of infinite extent and thickness  $2\bar{\mu}$  defined by the inequality  $|\mu| \leq \bar{\mu}$ . Let  $\mu^*$  be some control programme. Then

$$-\bar{\mu} \leq \mu^* + \xi \leq \bar{\mu} \quad (100)$$

Obviously from this it would be rather unreasonable to choose the programme (5) as follows:

$$\mu^* = \bar{\mu} \operatorname{sign} f(t) \quad (101)$$

Here  $f(t)$  is some switching function. If the number  $2\bar{\mu}$  actually exhausts the whole range of actuator displacement, then one is deliberately imposing very severe conditions on the actuator. Its displacements must satisfy the inequality

$$f_1(t) \leq \xi \leq f_2(t) \quad (102)$$

where  $f_1$  and  $f_2$  are given functions of  $t$  which assume values identically zero over whole sectors. Then it is always possible in the space  $A$  to designate some region  $A_\Sigma$  of initial values  $\eta_0$ , adjacent to the origin, which possesses the property: any sector  $R_0 = ||\eta_0||$  of  $A_\Sigma$  transforms into a sector  $R_T = ||\eta_T||$  of space  $B$  such that  $R_T > R_0$ . The greater  $T$ , the more intense is this

inequality. Obviously with the programme (101) the optimization problem loses its meaning. This is splendidly illustrated by the following example.

Let it be assumed that one is controlling a car on an absolutely rough horizontal plane.

The position of the car is defined by the coordinates  $x, y$  of its centre of gravity, the angle  $\psi$  between its plane of symmetry and the  $y$ -axis, and the angle  $\mu$  of rotation of its front wheels; the positive sense of the angles is taken as counterclockwise.

Suppose that it is required to choose a programme for optimal control of the car under the following boundary conditions: for  $t = 0$ , the coordinates of the car  $x_0 = y_0 = 0$ ,  $\psi_0 = 0$ ; while for  $t = T$  one has  $x_T < 0$ ,  $y_T > 0$ , and  $\psi = \psi_T > 0$ .

The solution is as follows: the wheels are rotated through an angle  $\mu^* = \bar{\mu}$ , which will ensure that the point  $x_T, y_T, \psi_T$  is reached. There is no requirement for switching of  $\mu$ .

Now suppose that the car has been replaced in its initial position  $x_0 = y_0 = 0$ , but with an angular error  $\Delta\psi(0) < 0$ . Obviously in this case the point  $x_T, y_T$  will never be reached. What is more, however small  $\Delta\psi(0)$  is, with the chosen control programme  $\mu^* = \bar{\mu}$  one can never be able to reduce the terminal control errors, which will be monotonically increasing with time.

Thus in formulating a statement of the optimal control problem one comes up against a *problem of compromise* in the separation of 'spheres of influence' such that allowance should be made both for the interests of control programme selection and those of optimal regulator design.

In many publications the subject is discussed in such a way as if this problem did not exist.

#### *Critique of the General Approach to Optimization. From Letters to the Author*

Letters to the author of this paper from Professors Balchen and Zadeh have drawn attention to the narrowness of the endeavour to optimize a system according to a single scalar criterion of the type (11).

In fact, when considering various vital situations in the broad sense of the word one intuitively tended to choose a plan of action under given conditions by assessing the situation from various points of view, which cannot be expressed by a common scalar criterion. Hence the idea is advanced in these letters of evaluating the response of a system according to a vector criterion. The author is in agreement with this criticism, but it is not possible here and now to give a precise statement to the problem of optimizing a system according to a vector criterion. This problem will form the subject of investigations in the immediate future.

Here only the speculation that it may perhaps not prove possible to formulate this problem correctly within the framework of the model (7) has been made. For this purpose it will possibly be necessary to consider as controlled plants what is now normally known as 'large systems'.

*The author has particular pleasure in acknowledging the help—in the form of books and papers, and by way of special manuscripts containing an exposition of various ideas and viewpoints on the problem of optimal control—which he has received from Professor A. I. Lur'e (Leningrad Polytechnical Institute), Dr. R. Bellman, Dr. Kalaba (Rand Corporation, California), Dr. R. Kulikowski (Warsaw Polytechnical Institute), Professor J. G. Balchen (Institute for Regelungstechnik, Trondheim), Professor L. Zahlen*

(University of California, Berkeley), Dr. T. Higgins (University of Wisconsin), Professor W. Locke (M.I.T.), Professor W. Oppelt (Technische Hochschule, Darmstadt) and Rotherbord Aris (University of Minnesota).

He expresses his sincere gratitude for their magnificent co-operation and for their unreserved efforts to facilitate the accomplishment of the task facing him.

### Bibliography

- <sup>1</sup> LIAPUNOV, A. M. *The General Problem of Stability of Motion*. 1950. Gostekhizdat
- <sup>2</sup> CHETAYEV, N. G. *The Stability of Motion*. 1956. Gostekhizdat
- <sup>3</sup> KRASOVSKIY, N. N. *Certain Problems in the Theory of Stability of Motion*. 1959. Moscow; Fizmatgiz
- <sup>4</sup> KOLMOGOROV, A. N. Analytical methods in probability theory. *Progr. math. Sci.*, Moscow, No. 5 (1938)
- <sup>5</sup> KHINCHIN, A. YA. Theory of the correlation of stationary stochastic processes. *Progr. math. Sci.*, Moscow, No. 5 (1938)
- <sup>6</sup> KOLMOGOROV, A. N. Interpolation and extrapolation of stationary random sequences. *Izv. Akad. Nauk SSSR, Ser. Mat.*, 5, No. 1 (1941)
- <sup>7</sup> WIENER, N. *Extrapolation, Interpolation and Smoothing of Stationary Time Series*. 1949. New York
- <sup>8</sup> ZADEH, L. A. and RAGAZZINI, I. R. Extension of Wiener's theory of prediction. *J. appl. Phys.* 21, No. 7 (1950)
- <sup>9</sup> BELLMAN, R. E. *Adaptive Control Processes. A Guided Tour*. 1961. Princeton University Press
- <sup>10</sup> PONTRYAGIN, L. S., BOLTYANSKIY, V. G., GAMKRELIDZE, R. V. and MISHCHENKO, YE. F. *The Mathematical Theory of Optimal Processes*. 1961. Moscow; Fizmatgiz
- <sup>11</sup> BELLMAN, R. E. and DREYFUS, S. F. *Applied Dynamic Programming*. 1962. Princeton University Press
- <sup>12</sup> *Optimization Technique with Applications to Aerospace Systems*. G. Leitmann. 1962. Academic Press
- <sup>13</sup> BELLMAN, R., GLICKSBERG, I. and GROSS, O. *Some Aspects of Mathematical Theory of Control Processes*. 1958. Santa Monica, California; Rand Corporation
- <sup>14</sup> BELLMAN, R. *Mathematical Optimization Techniques*. 1963. Berkeley; University of California Press
- <sup>15</sup> LANING, J. H. and BATTIN, R. H. *Random Processes in Automatic Control*. 1956. New York
- <sup>16</sup> PUGACHEV, V. S. The theory of random functions and its application to automatic control problems. 1957. Moscow; GITTL
- <sup>17</sup> KIPINIAK, W. *Dynamic Optimization and Control. A Variational Approach*. 1961. Cambridge (Mass.); M.I.T. Press
- <sup>18</sup> HOWARD, R. A. *Dynamic Programming and Markov Processes*. 1960. Cambridge (Mass.); M.I.T. Press
- <sup>19</sup> NEWTON, J. K., GOULD, L. A. and KAIZER, J. F. *Theory of Linear Servomechanisms*. 1961. Moscow; Fizmatgiz
- <sup>20</sup> MIELE ANGELO, *Flight Mechanics-I*. 1962. Massachusetts; Addison-Wesley
- <sup>21</sup> PETERSON, E. L. *Statistical Analysis and Optimization of Systems*. 1961. New York; Wiley
- <sup>22</sup> SHELDON S. L. CHANG *Synthesis of Optimum Control Systems*. 1961. New York; McGraw-Hill
- <sup>23</sup> BALCHEN, JENS G. Dynamic optimization of continuous processes. *Inst. for Regelungstechnik*. 1961. Trondheim
- <sup>24</sup> BUSHAW, D. W. Experimental towing tank. *Stevens Indic. Rep.* 469. 1953. Hoboken, N.J.
- <sup>25</sup> HOPKIN, A. M. A phase-plane approach to the compensation of saturating servomechanisms. *Trans. Amer. Inst. elect. Engrs* 70, No. 1 (1951) 631
- <sup>26</sup> FELDBAUM, A. A. Optimal processes in automatic control systems. *Automat. Telemekh.*, 14, No. 6 (1953)
- <sup>27</sup> LERNER, A. YA. On the limiting response speed of automatic control systems. *Automat. Telemekh.*, 15, No. 6 (1954)
- <sup>28</sup> KALMAN, R. E. Contributions to the theory of optimal control. *Symp. Int. de Ecuaciones Diferenciales Ordinarias*, 1951.
- <sup>29</sup> KALMAN, R. E. On a new approach to filtering and prediction problems. *J. Basic Engng Trans. Amer. Soc. mech. Engrs* 82 D (1960) 34
- <sup>30</sup> KALMAN, R. E. When is a linear control system optimal? *TR.* 63-5, March, 1963
- <sup>31</sup> KALMAN, R. E. On the general theory of control systems. *Proc. 1st Int. Congr. Automatic Control*, vol. 2. Published 1961 by the Academy of Sciences of the U.S.S.R. pp. 521-547
- <sup>32</sup> KALMAN, R. E. and BERTRAM, J. E. Control system analysis and design via the 'second method' of Liapunov. I. Continuous-time systems. *Trans. Amer. Soc. mech. Engrs*, 82 D, No. 2 (1960) 371
- <sup>33</sup> KALMAN, R. E. and BERTRAM, J. E. Control system analysis and design via the 'second method' of Liapunov. II. Discrete-time Systems. *Trans. Amer. Soc. mech. Engrs*, 82 D, No. 2 (1960) 394
- <sup>34</sup> MERRIAM, C. W. A class of optimum control systems. *J. Franklin Inst.* 267 (1959) 267
- <sup>35</sup> MERRIAM, C. W. Synthesis of adaptive controls. *M.I.T. Thesis*. 1958. Boston, Mass.
- <sup>36</sup> MERRIAM, C. W. Use of adaptive control systems. *Appl. Ind.*, No. 46 (1960) 506
- <sup>37</sup> LETOV, A. M. Analytical design of regulators. I. *Automat. Telemekh.* 21, No. 4 (1960)
- <sup>38</sup> LETOV, A. M. Analytical design of regulators. II. *Automat. Telemekh.* 21, No. 5 (1960)
- <sup>39</sup> LETOV, A. M. Analytical design of regulators. III. *Automat. Telemekh.* 21, No. 6 (1960)
- <sup>40</sup> LETOV, A. M. Analytical design of regulators. IV. *Automat. Telemekh.* 22, No. 4 (1961)
- <sup>41</sup> LETOV, A. M. Analytical design of regulators. V. Further development of the problem. *Automat. Telemekh.* 23, No. 11 (1962)
- <sup>42</sup> LUR'YE, A. I. A minimal quadratic criterion for the response of a control system. *Izv. Akad. Nauk SSSR, Tekh. Kibernetika* No. 4 (1963)
- <sup>43</sup> LITOVCHENKO, I. A. Contribution to the isoperimetric problem in the analytic design of an optimal regulator. *Automat. Telemekh.* 22, No. 12 (1961)
- <sup>44</sup> LITOVCHENKO, I. A. A problem in optimal control. *Automat. Telemekh.* 21, No. 8 (1960)
- <sup>45</sup> KURTSVEYL', YA. Contribution to the analytical design of regulators. *Automat. Telemekh.* 22, No. 6 (1961)
- <sup>46</sup> CHANG JEN-WEI A problem in the synthesis of optimal systems by the maximum principle. *Automat. Telemekh.* 22, No. 10 (1961)
- <sup>47</sup> CHANG JEN-WEI Contribution to the problem of optimal regulator synthesis in systems with delay. *Automat. Telemekh.* 23, No. 2 (1962)
- <sup>48</sup> AL'BREKHT, E. G. On the optimal stabilization of non-linear systems. *Appl. Math. Mech., Leningr.* 25, No. 5 (1961)
- <sup>49</sup> SALUKVADZE, M. YE. Analytical design of regulators. Continuously-acting perturbations. *Automat. Telemekh.* 22, No. 10 (1961)
- <sup>50</sup> SALUKVADZE, M. YE. On the analytical design of an optimal regulator with continuously-acting perturbations. *Automat. Telemekh.* 23, No. 6 (1962)
- <sup>51</sup> SALUKVADZE, M. YE. Contribution to the problem of optimal regulator synthesis in linear systems with delay that are subjected to continuously-acting perturbations. *Automat. Telemekh.* 23, No. 12 (1962)
- <sup>52</sup> KIRILOVA, L. S. A problem on the optimization of the terminal state of a control system. *Automat. Telemekh.* 23, No. 12 (1962)
- <sup>53</sup> KRASOVSKIY, N. N. and LIDSKIY, E. A. Analytical design of regulators in stochastic systems with limitations on the rate of change of the control action. *Appl. Math. Mech., Leningr.* 25, No. 3 (1961)

- <sup>54</sup> KRASOVSKIY, N. N. and LIDSKIY, E. A. Analytical design of regulators in systems with random properties. I. *Automat. Telemech.* 22, No. 9 (1961)
- <sup>55</sup> KRASOVSKIY, N. N. and LIDSKIY, E. A. Analytical design of regulators in systems with random properties. II. Equations for optimal control. Approximate method of solution. *Automat. Telemech.* 22, No. 10 (1961)
- <sup>56</sup> KRASOVSKIY, N. N. and LIDSKIY, E. A. Analytical design of regulators in systems with random properties. III. Optimal control in linear systems. Minimum R.M.S. error. *Automat. Telemech.* 22, No. 11 (1961)
- <sup>57</sup> KRASOVSKIY, N. N. On the choice of parameters for optimal stable systems. *Proc. 1st Int. Congr. Automatic Control*, 2, Published 1961 by the Academy of Sciences of the U.S.S.R., pp. 482 to 489
- <sup>58</sup> KRASOVSKIY, N. N. On R.M.S. optimal stabilization with random damping perturbations. *Appl. Math. Mech., Leningr.* 25 (1961) 806-817
- <sup>59</sup> KRASOVSKIY, N. N. On the analytical design of an optimal regulator in a system with time delay. *Appl. Math. Mech., Leningr.* 26, No. 1 (1962)
- <sup>60</sup> KRASOVSKIY, N. N. On a tracking problem. *Appl. Math. Mech., Leningr.* 26, No. 2 (1962)
- <sup>61</sup> KATS, I. YA. and KRASOVSKIY, N. N. On the stability of systems with random parameters. *Appl. Math. Mech., Leningr.* 24, No. 5 (1960)
- <sup>62</sup> LIDSKIY, E. A. On the stabilization of stochastic systems. *Appl. Math. Mech., Leningr.* 25, No. 5 (1961)
- <sup>63</sup> LIDSKIY, E. A. On the analytical design of regulators in systems with random properties. *Appl. Math. Mech., Leningr.* 26, No. 2 (1962)
- <sup>64</sup> LIDSKIY, E. A. Optimal control of systems with random properties. *Appl. Math. Mech., Leningr.* 27, No. 1 (1963)
- <sup>65</sup> LITVIN-SEDOY, M. Z. and SAVIN, A. B. On the synthesis of second-order automatic control systems with limited transient processes. *Vestn. Moskov. Gos. Univ., Matem. mekhan.* No. 5 (1961)
- <sup>66</sup> ZAYTSEV, A. G. Analytical design of systems reproducing a useful signal in the presence of noise. *Automat. Telemech.* 24, No. 2 (1963)
- <sup>67</sup> ZAYTSEV, A. G. Analytical design of optimal regulators with random perturbations. *Automat. Telemech.* 24, No. 4 (1963)
- <sup>68</sup> KRASOVSKIY, N. N. Contribution to optimal control theory. *Automat. Telemech.* 18, No. 11 (1957)
- <sup>69</sup> KRASOVSKIY, N. N. On a problem of optimal control. *Appl. Math. Mech., Leningr.* 21, No. 5 (1957)
- <sup>70</sup> KRASOVSKIY, N. N. Contribution to the problem of the existence of optimal trajectories. *Izv. Vuzov, Matem.* No. 6 (13) (1959)
- <sup>71</sup> KRASOVSKIY, N. N. Contribution to the sufficient conditions for optimization. *Appl. Math. Mech., Leningr.* 23, No. 3 (1959)
- <sup>72</sup> KRASOVSKIY, N. N. On a problem of optimal control in non-linear systems. *Appl. Math. Mech., Leningr.* 23, No. 2 (1959)
- <sup>73</sup> KRASOVSKIY, N. N. On optimal control in non-linear systems. *Izv. Vuzov, Matem.* No. 5 (12) (1959)
- <sup>74</sup> KRASOVSKIY, N. N. Contribution to optimal control theory. *Appl. Math. Mech., Leningr.* 23, No. 4 (1959)
- <sup>75</sup> KRASOVSKIY, N. N. On a method of constructing optimal trajectories. *Matem. Sbornik* 53, No. 2 (1960)
- <sup>76</sup> KRASOVSKIY, N. N. On optimal control with random perturbations. *Appl. Math. Mech., Leningr.* 24, No. 1 (1960)
- <sup>77</sup> KRASOVSKIY, N. N. On the approximate calculation of an optimal control by the direct method. *Appl. Math. Mech., Leningr.* 24, No. 2 (1960)
- <sup>78</sup> KRASOVSKIY, N. N. On the stabilization of unstable motions by additional forces with incomplete feedback. *Appl. Math. Mech., Leningr.* 25, No. 4 (1963)
- <sup>79</sup> POPOV, V. M. and KHALANAY, A. On a problem in the theory of optimal systems with delay. *Automat. Telemech.* 24, No. 2 (1963)
- <sup>80</sup> KRASOVSKIY, N. N. and LETOV, A. M. Contribution to the theory of the analytical design of regulators. *Automat. Telemech.* 23, No. 6 (1962)
- <sup>81</sup> KULIKOWSKI, R. Concerning the synthesis of the optimum non-linear control systems. *Bull. Acad. polon. sci. ser. sci., techn.* 7, No. 6 (1959)
- <sup>82</sup> KULIKOWSKI, R. On optimum control with constraints. *Bull. Acad. polon. sci., ser. sci. techn.* 7, No. 4 (1959)
- <sup>83</sup> KULIKOWSKI, R. On the synthesis of adaptive systems. *Bull. Acad. polon. sci., ser. sci. techn.* 7, No. 12 (1959)
- <sup>84</sup> KULIKOWSKI, R. Synthesis of a class of optimum control systems. *Bull. Acad. sci. polon., ser. sci. techn.* 7, No. 11 (1959) 663
- <sup>85</sup> KULIKOWSKI, R. Synthesis of optimum control systems with area-bounded control signal. *Bull. Acad. polon. sci., ser. sci. techn.* 8, No. 4 (1960) 179
- <sup>86</sup> MESAROVIC, M. On the existence and uniqueness of the optimal multivariable system synthesis. *Inst. Radio Engrs Trans. on Automatic Control*
- <sup>87</sup> LA SALLE, J. P. Time optimal control. *Proc. Nat. Acad. Sci., Wash.* 45, No. 4 (1959) 573
- <sup>88</sup> LEE, E. B. and MARKUS, L. Optimal control for non-linear processes. *Archive for Nat. Mech. and Anal.* 8 (1961)
- <sup>88a</sup> ZUBOV, V. I. *Oscillations in Non-linear Systems*. 1962. Sudpromgiz
- <sup>89</sup> OKHOTIMSKIY, D. YE. and ENEYEV, T. M. Some variational problems connected with the launching of an artificial Earth satellite. *Adv. Phys. Sci., Moscow* 63, No. 1a (1957)
- <sup>90</sup> KIRILLOVA, F. M. On a limit transition in the solution of an optimal control problem. *Appl. Math. Mech., Leningr.* 24, No. 2 (1960)
- <sup>91</sup> KIRILLOVA, F. M. Contribution to the problem of analytical regulator design. *Appl. Math. Mech., Leningr.* 25, No. 3 (1961)
- <sup>92</sup> KIRILLOVA, F. M. Contribution to the problem of the existence of optimal trajectories for non-linear systems. *Izv. Vuzov, Matem.* No. 2 (21) (1961)
- <sup>93</sup> KIRILLOVA, F. M. Some questions in optimal control theory. *Izv. Vuzov, Matem.* No. 3 (1962)
- <sup>94</sup> KIRILLOVA, F. M. Some questions in optimal control theory. *Progr. Math. Sci., Moscow* 17, No. 1 (1962)
- <sup>95</sup> KIRILLOVA, F. M. On the continuous dependence of the solution of an optimal control problem on the initial data and parameters. *Progr. Math. Sci., Moscow* 17, No. 4 (1962)
- <sup>96</sup> KIRILLOVA, F. M. On the correctness of the statement of an optimal control problem. *Izv. Vuzov, Matem.* No. 4 (5) (1959)
- <sup>97</sup> FILIPPOV, A. F. On certain questions in optimal control theory. *Vestn. Moskov. Gos. Univ., Ser. Matem. mekhan.* No. 2 (1959)
- <sup>98</sup> BOLTYANSKIY, V. G. Sufficient conditions for optimization. *C. R. Acad. Sci. U.R.S.S.* 140, No. 5 (1961)
- <sup>99</sup> GAMKRELIDZE, R. V. Contribution to the general theory of optimal processes. *C. R. Acad. Sci. U.R.S.S.* 123, No. 2 (1958)
- <sup>100</sup> GAMKRELIDZE, R. V. Contribution to the theory of optimal processes in linear systems. *C. R. Acad. Sci. U.R.S.S.* 116, No. 1 (1957)
- <sup>101</sup> GAMKRELIDZE, R. V. Optimal control processes with limited phase coordinates. *Bull. Acad. Sci. U.R.S.S. Ser. Math.* 24, No. 3 (1960)
- <sup>102</sup> TROITSKIY, V. A. Variational problems in the optimization of control processes for equations with discontinuous R.H.S. *Appl. Math. Mech., Leningr.* 26, No. 2 (1962)
- <sup>103</sup> TROITSKIY, V. A. Variational problems in the optimization of control processes in systems with limited coordinates. *Appl. Math. Mech., Leningr.* 26, No. 3 (1962)
- <sup>104</sup> TROITSKIY, V. A. The Mayer-Boltz problem in calculus of variations and the theory of optimal systems. *Appl. Math. Mech., Leningr.* 25, No. 4 (1961)

- <sup>105</sup> TROITSKIY, V. A. On variational problems of control process optimization. *Appl. Math. Mech., Leningr.* 26, No. 1 (1962)
- <sup>106</sup> TROITSKIY, V. A. Variational problems of control process optimization using functionals depending on the intermediate values of the coordinates. *Appl. Math. Mech., Leningr.* 26, No. 6 (1962)
- <sup>107</sup> TROITSKIY, V. A. Variational problems of control process optimization using functionals depending on the intermediate values of the coordinates. *C. R. Acad. Sci. U.R.S.S.* 149, No. 2 (1963)
- <sup>108</sup> KROTOV, V. F. On discontinuous solutions in variational problems. *Izv. Vuzov, Matem.* No. 2 (21) (1961)
- <sup>109</sup> KROTOV, V. F. On the absolute minimum of functionals on the basis of functions with limited derivative. *C. R. Acad. Sci. U.R.S.S.* 140, No. 3 (1961)
- <sup>110</sup> KROTOV, V. F. The basic problem of the calculus of variations in relation to discontinuous functions. *C. R. Acad. Sci. U.R.S.S.* 137, No. 1 (1961)
- <sup>111</sup> KROTOV, V. F. Discontinuous solutions of variational problems. *Izv. Vuzov, Matem.* No. 5 (18) (1960)
- <sup>112</sup> KROTOV, V. F. Methods of solving variational problems based on sufficient conditions for absolute minimum. I. *Automat. Telemech.* 23, No. 12 (1962)
- <sup>113</sup> ROZONOER, L. I. Integral response indicators in automatic control theory, and assessment of the behaviour of the function from the known value of the functional. *Bull. Acad. Sci. U.R.S.S., Otd. Tekn. Nauk. Energ. i Automat.*, No. 4 (1959)
- <sup>114</sup> ROZONOER, L. I. On sufficient conditions for optimization. *C. R. Acad. Sci. U.R.S.S.* 127, No. 3 (1959)
- <sup>115</sup> ROZONOER, L. I. Pontryagin's maximum principle in optimal system theory. I. *Automat. Telemech.* 20, No. 10 (1959)
- <sup>116</sup> ROZONOER, L. I. Pontryagin's maximum principle in optimal system theory. II. *Automat. Telemech.* 20, No. 11 (1959)
- <sup>117</sup> ROZONOER, L. I. Pontryagin's maximum principle in optimal system theory. III. *Automat. Telemech.* 20, No. 12 (1959)
- <sup>118</sup> ROYTENBERG, YA. N. Some problems in dynamic programming theory. *Appl. Math. Mech., Leningr.* 23, No. 4 (1959)
- <sup>119</sup> ROYTENBERG, YA. N. Some problems in the theory of dynamic programming for non-linear systems. *Appl. Math. Mech., Leningr.* 26, No. 3 (1962)
- <sup>120</sup> STRATONOVICH, R. L. Contribution to optimal control theory. Sufficient coordinates. *Automat. Telemech.* 23, No. 7 (1962)
- <sup>121</sup> STRATONOVICH, R. L. On the optimal observation of derangement of a production process. *Vestn. Moskov. Gos. Univ., Ser. Matem. mekh.*, No. 2 (1962)
- <sup>122</sup> STRATONOVICH, R. L. Conditional Markov processes in problems of mathematical statistics and dynamic programming. *C. R. Acad. Sci. U.R.S.S.* 140, No. 4 (1961)
- <sup>123</sup> STRATONOVICH, R. L. Contribution to optimal control theory. Asymptotic method of solving the diffusion alternative equation. *Automat. Telemech.* 23, No. 11 (1962)
- <sup>124</sup> KUKHTENKO, V. I. Contribution to the calculation of correction networks for automatic control systems from the criterion of minimum R.M.S. error.
- <sup>125</sup> SOLODOVNIKOV, V. V. and MATVEYEV, P. S. Synthesis of corrective devices for servomechanisms in presence of noise with given requirements for dynamic accuracy. *Autom. Telemech.* 16, No. 3 (1955)
- <sup>126</sup> NAUMOV, B. N. Synthesis of non-linear automatic control systems. In book 'Results of science. Vol. 1. Problems of non-linear automatic control system theory.' *Bull. Acad. Sci. U.R.S.S.* (1957) 105-132
- <sup>127</sup> SOLODOVNIKOV, V. V. Statistical dynamics of linear automatic control systems.
- <sup>128</sup> FEL'DBAUM, A. A. Theory of dual control. I, II. *Automat. Telemech.* 21, Nos. 9, 11 (1960)
- <sup>129</sup> FEL'DBAUM, A. A. Theory of dual control. III. *Automat. Telemech.* 22, No. 1 (1961)
- <sup>130</sup> FEL'DBAUM, A. A. Theory of dual control. IV. *Automat. Telemech.* 22, No. 2 (1961)
- <sup>131</sup> BELLMAN, R. Dynamic Programming and Stochastic Control Processes. *Inform. Control* 1 (1958) 228
- <sup>132</sup> BELLMAN, R. and KALABA, R. Dynamic programming and adaptive processes—A mathematical foundation. *Inst. Radio Engrs Trans. Auto. Control* AC-5 (1960) 5
- <sup>133</sup> BELLMAN, R. and KALABA, R. Functional equations in adaptive processes and random transmission. *Inst. Radio Engrs Trans. Circuit Theory* CT-6 (1959) 271
- <sup>134</sup> BELLMAN, R. and KALABA, R. Dynamic Programming applied to control processes governed by general functional equations. *Proc. Nat. Acad. Sci., Wash.* 48 (1962) 1735
- <sup>135</sup> BELLMAN, R. A mathematical formulation process of adaptive type. *IVth Berkeley Symp. on Math. Stat. and Probability* 1 (1961) 37. Berkeley; Univ. of California, Press
- <sup>136</sup> BELLMAN, R. On the foundations of a theory of stochastic variational processes. *Proc. Techn. Symp. in Appl. Math. Soc.* 1 (1961). Providence R.
- <sup>137</sup> BELLMAN, R. Directions of mathematical research in nonlinear circuit theory. *Inst. Radio Engrs Trans. Circuit Theory* CT-7 (1960) 542
- <sup>138</sup> BELLMAN, R. New direction of research in the theory of differential equations. Nonlinear differential equations. *Nonlinear Differential Equations and Nonlinear Mechanics* (1963) 121-134
- <sup>139</sup> BELLMAN, R. Functional equations in adaptive processes and random transmission. *Trans. Int. Symp. on Circuit and Information Theory* (1959) 271-282
- <sup>140</sup> BELLMAN, R. and BROOK, R. On the concepts of problem solving. *Amer. Math. Mon.* 67 (1960) 119
- <sup>141</sup> ADORNO, D. S. The asymptotic theory of control systems-1. Stochastic and deterministic processes. *Jet Propulsion Lab., Technical Release* 34-73 (June 30, 1960)
- <sup>142</sup> AOKI, M. Minimising integrals of absolute deviation in linear control systems. *Appl. and Ind.*, No. 61 (1962) 125
- <sup>143</sup> AOKI, M. Dynamic programming and numerical experiments as applied to adaptive control systems. *Ph. D. Thesis, Dept. of Engng* (Nov. 1959). Los Angeles; Univ. of California
- <sup>144</sup> AOKI, M. Dynamic programming approach to a final value control system with a random variable having an unknown distribution function. *Inst. Radio Engrs Trans. Auto. Control* AC-5, No. 4 (1960) 270
- <sup>145</sup> AOKI, M. On optimal and suboptimal policies in the choice of control forces for final-value systems. *Inst. Radio Engrs Trans. Auto. Control* AC-5, No. 3 (1960) 171
- <sup>146</sup> AOKI, M. Stochastic time optimal-control systems. *Appl. and Ind.* No. 54 (1961) 41
- <sup>147</sup> BARBASHIN, YE. A. On the assessment of R.M.S. deviation from a given trajectory. *Automat. Telemech.* 21, No. 7 (1960)
- <sup>148</sup> BARBASHIN, YE. A. On the assessment of the maximum deviation from a given trajectory. *Automat. Telemech.* 21, No. 10 (1960)
- <sup>149</sup> BARBASHIN, YE. A. On a problem of dynamic programming theory. *Appl. Math. Mech., Leningr.* 24, No. 6 (1960)
- <sup>150</sup> BARBASHIN, YE. A. On the approximate establishment of motion along a given trajectory. *Automat. Telemech.* 22, No. 6 (1961)
- <sup>151</sup> BUTKOVSKIY, A. G. Optimal processes in systems with distributed parameters. *Automat. Telemech.* 22, No. 1 (1961)

- <sup>152</sup> BUTKOVSKIY, A. G. The maximum principle for optimal systems with distributed parameters. *Automat. Telemekh.* 22, No. 10 (1961)
- <sup>153</sup> TSYPKIN, YA. Z. On optimal processes in pulsed automatic systems. *C. R. Acad. Sci. U.R.S.S.* 134, No. 2 (1960)
- <sup>154</sup> TSYPKIN, YA. Z. Optimal processes in pulsed automatic systems. *Bull. Acad. Sci. U.R.S.S. Otd. Tekh. Nauk Energ. i Automat.* No. 4 (1960)
- <sup>155</sup> CH'EN HSÜEH-SEN *Technical Cybernetics*. 1956. *Izd. Inostr. Lit.*
- <sup>156</sup> RUTHERFORD, A. *The Optimal Design of Chemical Reactors*. 1961. New York; Academic Press
- <sup>157\*</sup> Joint Automatic Control Conference 1962
- <sup>158\*</sup> *J. Basic Engng* June (1960), March (1962)
- <sup>159\*</sup> GIBSON, J. E. Proc. dynamic programming workshop. 1961
- <sup>160\*</sup> Proc. 1st Congr. Int. Federation on Automatic Control. 1961. Moscow

\* Papers devoted to automatic control problems.



# Approximation Methods in Optimal and Adaptive Control

J. H. WESTCOTT, J. J. FLORENTIN and J. D. PEARSON

## Summary

The mathematical equations for optimal and adaptive control are now known. However, comprehensive methods for their numerical solution are not available. This paper gives a discussion of approximate computer methods of finding numerical solutions to suitable engineering accuracy. The importance of setting up the problem in a suitable form is stressed, and illustrated with an example. Three explicit approximation techniques are described, one a trajectory method, and the others function space methods. For each technique a worked example is given.

## Sommaire

Les équations mathématiques des systèmes d'optimisation et d'adaptation automatique sont connues. Ce qui l'est moins, ce sont les méthodes numériques permettant d'en déterminer les solutions. Le rapport décrit quelques méthodes d'approximation au moyen de calculateurs conduisant à des solutions numériques dont l'exactitude est suffisante pour les besoins de la pratique. Il met en évidence la nécessité de mettre les données du problème sous une forme appropriée et illustre cette nécessité par un exemple. Trois méthodes d'approximation explicites sont décrites, l'une se rapportant à une trajectoire à une dimension, les deux autres à des fonctions à plusieurs dimensions; chacune de ces méthodes est illustrée par un exemple.

## Zusammenfassung

Mathematische Gleichungen für optimale und selbststeuende Regelungssysteme sind bekannt, jedoch gibt es noch keine umfassenden Methoden für deren numerische Lösung. Dieser Aufsatz behandelt einige auf dem Rechner durchführbare Näherungsverfahren, die zu numerischen Lösungen führen, die für praktische Zwecke ausreichen. Besonders wird darauf hingewiesen — und anhand eines Beispiels erklärt — wie wichtig es ist, das Problem richtig zu formulieren. Drei explizite Näherungsverfahren werden besprochen, von denen eines auf der Darstellung durch Trajektorien und die anderen auf der Darstellung im Funktionenraum beruhen. Für jede dieser Methoden wird ein durchgerechnetes Beispiel angeführt.

## Introduction

Both optimal and adaptive control problems can now be treated by the same decision theory approach<sup>1</sup>. Typical practical problems can be formulated in the required mathematical terms, but at present there is still difficulty in determining actual numerical solutions to problems of realistic size and complexity. It seems likely that a variety of approximate computation techniques will be developed, each with a restricted range of application.

Approximation is necessary due to the very extensive calculation called for, using multi-stage decision methods. These become sufficiently time-consuming, even when performed at the fast speeds of modern digital computers, for abbreviation to be necessary. Approximation becomes attractive due to the limited time allowable on the time scale of the dynamic process it is desired to control.

Approximation may be attempted either in the setting up of a particular problem or during the numerical process of solution. Whilst approximations in the setting up procedure are somewhat difficult to treat analytically they are likely to be important in practical applications of decision theory, and a worked example is presented to demonstrate some aspects of this. In numerical techniques two broad approaches can be distinguished—trajectory and function space methods. Certain approximation techniques in both these classes are discussed and illustrated with simple examples.

## Review of Basic Equations

To introduce the approach and notation, a brief review of the basic equations will be given. More complete derivations can be found<sup>1, 2</sup>. The complete system including inputs must first be described by a set of state coordinates  $\mathbf{x}(t)$ . These are an absolute description of the system at the time instant  $t$ . The state coordinates may be obvious physical quantities such as position or velocity, or may be statistical quantities such as the mean and the variance. The motion of the system is conveniently described by a set of first order vector differential equations:

$$\mathbf{x}' = \mathbf{A}(\mathbf{x}, \mathbf{u}, t) \quad (1)$$

where  $\mathbf{u}$  is a vector of control variables. For reasons of simplicity no random components are included.

An optimal control function  $\mathbf{u}(t)$  is to be found which maximizes or minimizes a given performance index, subject to constraints. Typically the performance index will be a path integral over a defined time interval  $t, T$  with the system starting at a given position  $\mathbf{x}$ . Here the performance index will be defined as

$$f(\mathbf{x}(t), t) = \min_{\mathbf{u}} \int_t^T L(\mathbf{x}, \mathbf{u}, \lambda) d\lambda \quad (2)$$

Isolating a small part of the time interval  $t, T$ , it can be seen<sup>8</sup> that the following problem is equivalent to (2)

$$f(\mathbf{x}(t), t) = \min_{\mathbf{u}} \left[ \int_t^{t+\Delta} L(\mathbf{x}, \mathbf{u}, \lambda) d\lambda + f(\mathbf{x}(t+\Delta), t+\Delta) \right] \quad (3)$$

Using a Taylor series expansion for the third term, a partial differential equation for the performance index can be found in the form

$$\frac{\partial f}{\partial t} = - \min_{\mathbf{u}} \left[ \sum_i^n A_i \frac{\partial f}{\partial x_i} + L \right] \quad (4)$$

The solution of this partial differential equation reveals the performance index as a function of the initial position and time interval. The minimizing control law  $\mathbf{u}(t)$  can be found as a function of the partial derivatives of  $f$ .

With appropriate modifications analogous equations can be developed for discrete time systems and for those containing

random elements. When there are random elements the performance index must be changed to an ensemble average of path integrals<sup>3</sup>.

For deterministic systems eqn (3) can be expressed in an alternative form by taking its characteristics. Define the function  $H$  as

$$H\left(x, \frac{\partial f}{\partial x}, t\right) = \sum_1^n A_i \frac{\partial f}{\partial x_i} + L \quad (5)$$

Employing the normal theory of partial differential equations, a new set of variables,  $p_i$ , is defined with a vector  $p = \partial f / \partial x$ . The set of equations for the characteristics is then

$$\dot{x} = \frac{\partial H}{\partial p} \quad \dot{p} = -\frac{\partial H}{\partial x} \quad (6)$$

To solve these equations the boundary conditions of the  $p$  or co-state variables are required. As derived by Rozonoer<sup>4</sup>, the problem can correspond either to a fixed end point variational problem in which

$$x(T) \text{ known} \quad p(T) \text{ unknown} \quad (7)$$

or to a free end point variational problem in which

$$x(T) \text{ may be chosen} \quad p(T) = 0 \quad (8)$$

In either case the problem reduces to a two-point boundary value one.

### Computational Methods

Eqns (3) and (6) give rise to two families of computational techniques. The first is of wide application, it computes  $f(x, t)$  over the  $x$  space at successive time intervals  $t_r$ . It will be termed the function space method. A predominant difficulty in the function space approach is that of storing a multi-dimensional function in a digital computer. The obvious methods become prohibitive, since for a three-dimensional function with a hundred points in each dimension a storage space of  $10^6$  words would be necessary.

The second method applies only to deterministic systems and is termed the trajectory method. Only half of the boundary conditions at each end of the trajectory are known and the usual difficulties are encountered.

### Setting Up Problems

In practical situations the problem of control is usually not completely defined. In many cases it is possible to complete the specification by the selection of constraints which enable a

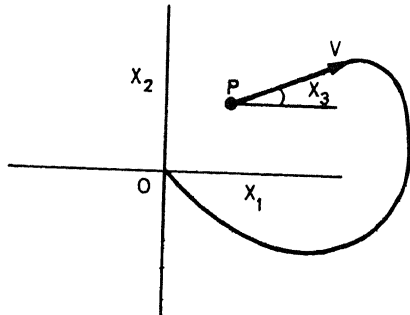


Figure 1

simpler solution to be achieved than would otherwise be possible. The factors which can most usefully and easily be varied are the mathematical form of the performance index, the precise specification of the control constraints, and the selection of continuous or discrete time working.

To illustrate the effects of varying the formulation consider a typical non-linear problem. A vehicle, moving with constant velocity  $v$  in the  $x_1, x_2$  phase plane, is to be guided from an initial point  $P$  to a final target  $O$ . Guidance is affected by adjusting the rate of turn of the vehicle. For practical reasons the maximum rate of turn is constrained. In accordance with Figure 1 the dynamic equations of the problem are:

$$\begin{aligned} \dot{x}_1 &= v \cos x_3 \\ \dot{x}_2 &= v \sin x_3 \\ \dot{x}_3 &= u \end{aligned} \quad (9)$$

where  $x_3$  is the angle the velocity vector makes with the  $x_1$  axis.

### Minimum Time Trajectories

Take the performance index to be

$$f(x, t) = \min_u \left[ \int_t^T h \cdot dt \right]$$

with the definition  $h = 0$  in a region surrounding the origin and  $h = 1$  elsewhere.

Following the usual formulation

$$H = \min_u [h + p_1 v \cos x_3 + p_2 v \sin x_3 + p_3 u]$$

leading to the optimal  $u$  being

$$u = -\text{sign}(p_3) \quad |u| \leq 2$$

The trajectory equations follow from the characteristics of  $H$  and reveal that since

$$p'_1 = p'_2 = 0$$

$$p'_3 = p_1 v \sin x_3 - p_2 v \cos x_3 \quad (10)$$

the optimal trajectories are composed of straight lines and circles depending on the boundary conditions of  $p_1$  and  $p_2$ .

As a physical consideration the additional definition of

$$\text{sign}(0) = 0$$

minimizes the number of switchings in  $u$  and the control is one commonly termed 'bang-bang'. For the chosen initial point  $P$  the optimal trajectories are at  $A$  in Figures 2 and 3. The vehicle reached the origin  $O$  in 3.24 sec.

### Minimum Energy Trajectories

If the performance index is taken to be

$$f(x, t) = \min_u \left[ \int_t^T \left( h + \frac{u^2}{2v} \right) d\lambda \right]$$

with  $v$  as a yet unchosen Lagrange multiplier, the optimum  $u$  is  $u = -v p_3$ . The trajectory equations are unchanged. At the terminal point  $O$ , the  $x_3$  boundary is unspecified and consequently

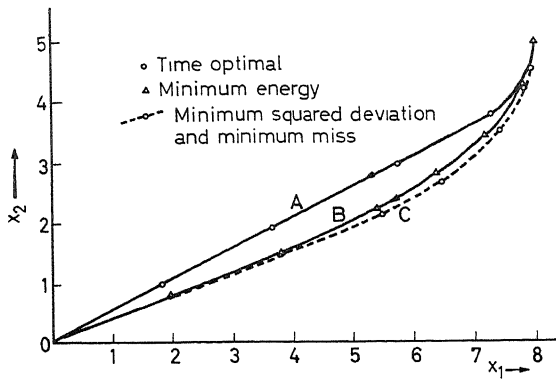


Figure 2.

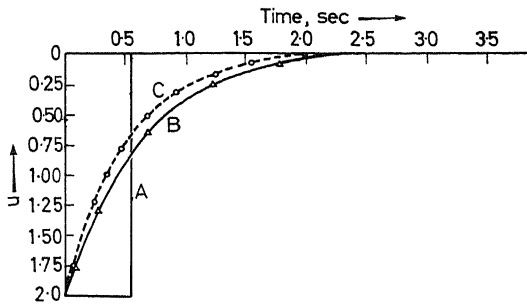


Figure 3. Control signals

$p_3$  and  $u$  are zero. To the first order of approximations along the trajectory through the origin

$$df = \frac{dx_1}{v} \cos x_3 + \frac{dx_2}{v} \sin x_3$$

and hence

$$\frac{\partial f}{\partial x_1} = p_1 = \frac{\cos x_3}{v} \quad \frac{\partial f}{\partial x_2} = p_2 = \frac{\sin x_3}{v}$$

The problem of solving the trajectory equations reduces to the selection of a terminal  $x_3$  and adjusting it until the initial boundary values are satisfied. For the given values of  $v = 3$  the optimal trajectories are shown at B in Figures 2 and 3. The system reached the origin from P in 3.27 sec.

#### Minimum Squared Deviation and Control Energy

The performance index is taken to be

$$f(x, t) = \min_u \int_t^T \frac{1}{2} \left( x_1^2 + x_2^2 + \frac{u^2}{v} \right) d\lambda$$

from which the optimal control is unchanged at  $u = -v p_3$ . However, the trajectory equations are now

$$\begin{aligned} p_1' &= -x_1 \\ p_2' &= -x_2 \\ p_3' &= p_1 v \sin x_3 - p_2 v \cos x_3 \end{aligned}$$

with free boundary value conditions.

The computational problem is not attractive in this case because at the terminal point,  $p_3'$  is very nearly always zero and tends to be influenced by rounding errors. The optimal trajectories are shown at C in Figures 2 and 3 and the optimal time to reach the origin was 3.29 sec.

#### Minimum Squared Miss Distance

The performance index is taken to be

$$f(x, t) = \min_u \left[ \int_t^T \frac{u^2}{2v} d\lambda \right] + \frac{1}{2} x_1^2(T) + \frac{1}{2} x_2^2(T)$$

resulting in the same control signal  $u = -v p_3$ . The trajectory equations are those of (10) and (9) and the boundary conditions are clearly  $p_1(T) = x_1$ ,  $p_2(T) = x_2$ ,  $p_3(T) = 0$ . Since the target at 0 is the origin these values closely approximate those for the minimum energy trajectory presented earlier in this paper. The optimal solutions are those for C in Figures 2 and 3 obtained for  $v = 100$ , and the minimum time achieved was 3.27 sec.

#### Summary of Formulations

Comparing the energy used in each of the four schemes reveals that apart from the minimum time trajectory there was little difference. The minimum time control scheme used 2.16 units as opposed to a minimum energy formulation using 1.23 units.

The insensitivity of the solution to the formulation suggests that the performance index producing the most tractable set of equations should be employed. In this example the minimum time solution is a geometric exercise, whilst the formulation of the minimum squared deviation and control energy generated a rather troublesome set of equations from the numerical point of view. Clearly, in general, there is room for experiment within the constraints of the problem.

#### Approximation Method for Boundary Value Problems

Since the trajectory methods lead to two point boundary value problems, methods for dealing with these are receiving a great deal of attention. Boundary value problems are far older than numerical variational methods, and many well established techniques exist in numerical practice. It is essential to distinguish between the time available for calculations which are of a design nature, and those that are undertaken during the control of the physical process. Real time work places a heavy penalty on inefficient computational techniques, and boundary value methods tend to be among the most inefficient.

#### Refining Techniques

Free boundary value problems of the type of eqn (8) lend themselves to reverse time computation from the guessed target position at time  $T$ . However, owing to the extreme sensitivity of the solutions and to some troublesome numerical details, it proves far simpler to work forwards from the initial time  $t$ . Consider the boundary value problem of eqn (8) and assume that it is possible to proceed by guessing the initial value of  $p(t)$  and adjusting this according to the error at the terminus. Represent the solution (6) for the terminal value of  $p$  by

$$p(T) = \Psi[p(t), x(t); t, T] \quad (11)$$

If the initial value is correctly chosen then, clearly,  $\Psi$  will be zero. Furthermore, if  $T$  is extended, the change in the correct  $p(t)$  will satisfy the differential equation

$$\left[ \frac{\partial \Psi_i}{\partial p_j(t)} \right] \cdot \frac{dp_j(t)}{dT} + \frac{\partial \Psi}{\partial T} = 0 \quad (12)$$

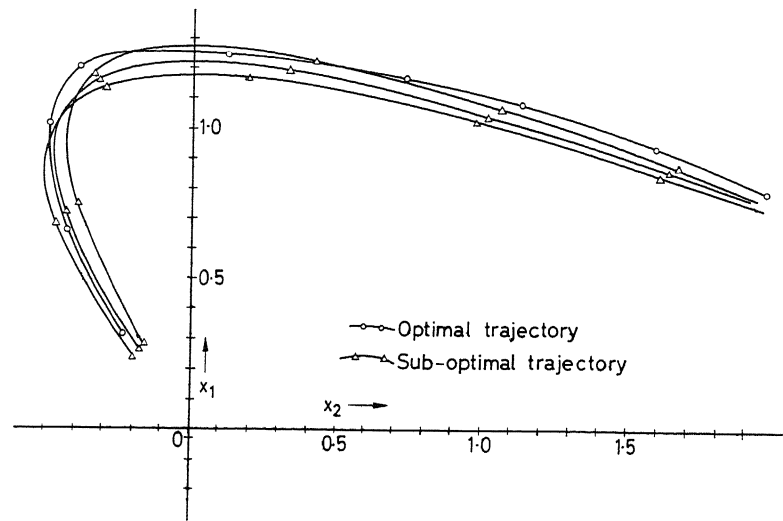


Figure 4. Varying performance index

However, although boundary values for these equations are known the matrix  $[\partial \Psi_i / \partial p_j]$  is not. In general  $p(T)$  will be non-zero at time  $T$  and the magnitude of its error can be evaluated by some arbitrary definite function  $n[p(t)]$  of the terminal boundary values. Small perturbations in the initial value  $p(t)$  will affect the value of  $n$  according to the expansion

$$n(p + \Delta p) = n(p) + \sum \frac{\partial n}{\partial p_i} \cdot \Delta p_i + O(\Delta^2 p_i) \quad (13)$$

However, since  $n$  can be defined to have a minimum when the boundary value is satisfied, a correction scheme can be found by differentiation with respect to one of the  $p_i$ :

$$\Delta p_i = - \frac{\partial n}{\partial p_i} / \frac{\partial^2 n}{\partial p_i^2} \quad (14)$$

and the corrected value of  $p_i$  will be  $p_i + \Delta p_i$ . The drawback of this simple approach is that the instability of the trajectory equations usually makes any function  $n$  extremely large and grossly sensitive to perturbations. It is, however, very simple and easy to code into a routine which could deal with all cases likely to arise in practice.

One way of overcoming this is to reduce substantially the time interval  $T-t$ . Clearly if this interval were zero the boundary values are known to be zero and consequently for a small interval a good estimate is available. Thus the procedure is to start with a small value  $T-t$ , perturb each of the  $p$  co-ordinates in turn, applying eqn (14) to reduce  $p(t')$  to zero,  $t'$  being the temporary value of  $T$ . After one cycle of the perturbations it is essential to rotate the axis of the perturbation coordinates to make them lie along the direction of  $\text{grad } n$ . Having solved the first stage over the reduced interval, it can be extended and the process repeated. As the next guess for the boundary values either the previous value can be used or it can be up-dated with a crude solution of eqn (12). The matrix  $[\partial \Psi_i / \partial p_i]$  is now available from the results of the previous perturbations and is invertible.

Proceeding in this way, the optimal trajectory can be progressively extended until it covers the given time interval. However, it may happen that some physical condition is satisfied

before the full time, i.e. the target region, is entered, and then the computation can be terminated earlier with consequent economy. This procedure has the advantage that during real time computation a 'part-time' optimal solution is always available for the current position and this can be used as an approximation while its up-dated value is refined. It has the disadvantage that errors in the numerical integration tend to disguise the minimum sought for, and these increase with  $T-t$ .

#### A Boundary Value Example

The method can be illustrated by an example whose non-linearities cause difficulties with the more conventional techniques. Consider a dynamic system of the pendulum type

$$x'' - (x')^2 + x = u \quad (15)$$

Select the control law  $u$  to minimize the performance index

$$f(x, t, T) = \frac{1}{2} \int_t^T (x_1^2 + x_2^2 + u^2) dt \quad (16)$$

The dynamic system is controllable only in the absence of limits on  $u$ . Transforming to the usual phase coordinates and forming the optimal equations yields the set

$$\begin{aligned} x'_1 &= x_2 \\ x'_2 &= x_2^2 - x_1 - p_2 \\ p'_1 &= -x_1 + p_2 \\ p'_2 &= -x_2 - p_1 - 2x_2 p_2 \end{aligned} \quad (17)$$

If the end point is considered to be free, then the boundary values are  $p(T)$  zero. Direct computation of the trajectories backwards from the terminal region surrounding the origin reveals an optimum phase portrait of Figure 5. It shows that it is difficult to reach the chosen initial point ( $x_1 = 1, x_2 = 0$ ) because of the sensitivity of the trajectories.

At the origin the non-linear system behaves as the linear one:

$$x' = A \cdot x$$

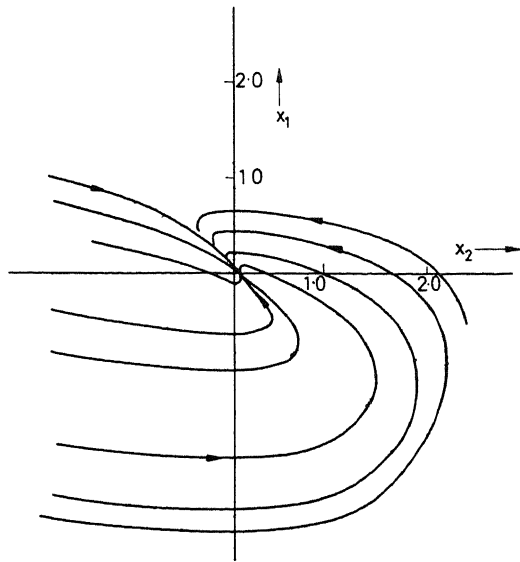


Figure 5. Optimal phase plane

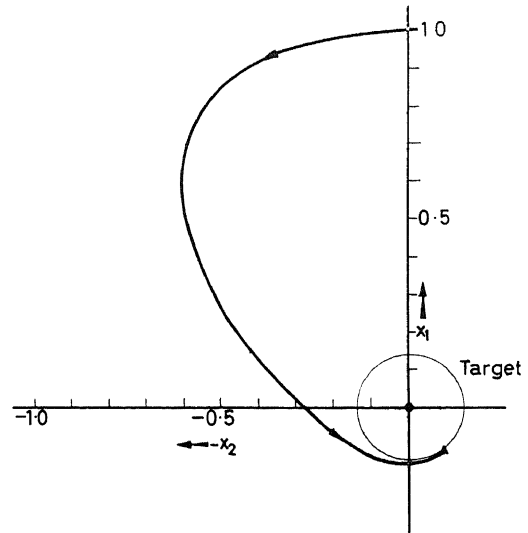


Figure 6. Optimal trajectory

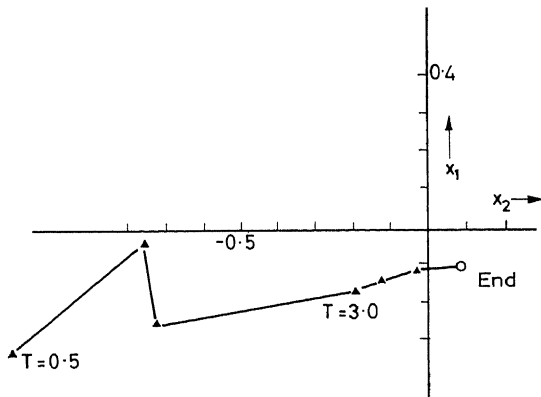


Figure 7. Variation of end point trajectory

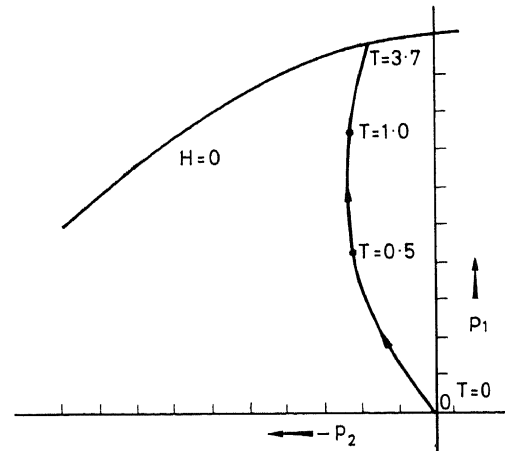


Figure 8. Variation of optimal boundary values

The solution of this system as  $t$  tends to infinity, is dominated by that of the largest positive eigenvalue of  $A$ , and lies parallel to the corresponding eigenvector. Thus all solutions tend to the one eigenvector through the origin and this implies extreme sensitivity when working in reverse time scale from the origin.

Using the technique described, which has been programmed into a series of short routines for the Ferranti Mercury computer, a trajectory can be found which satisfies any initial conditions and boundary values. Figure 6 gives the final trajectory and also the variation of the end point  $x(T)$  with the time interval  $T-t$  given in Figure 7. Figure 8 shows how the computed finite time boundary values tend to the Hamiltonian surface  $H = 0$  satisfied for the infinite interval.

### Function Space Approximations

Function space methods are an alternative to trajectory methods in deterministic systems, but are the only approach in systems containing random components. In some systems the algebraic form of  $f(x, t)$  is known in advance and the computational formulae may then be put into discrete or continuous

time form, whichever is most convenient. However, when the form of  $f(x, t)$  is unknown, only discrete time computing formulae are feasible. This is usually the situation in adaptive systems.

### A Function Space Approximation for an Adaptive System

To illustrate the use of function space approximations consider the following example. A simple regulator contains a fixed unknown gain  $\alpha$ , in the control path (Figure 9). The system is disturbed by a noise  $\{\xi_n\}$  which is an independent Gaussian sequence with zero mean and variance  $\sigma$ . In order to set up the problem to lead to discrete time computing formulae it is assumed that the control is changed only at unit time

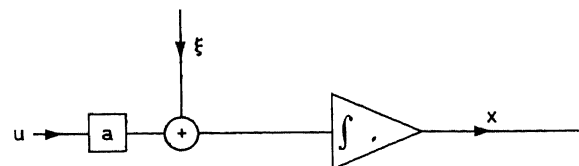


Figure 9. Regulator with unknown gain

intervals, when  $x$  is also observed. The dynamic equation for the regulator is

$$x_n = x_{n-1} + \alpha u_{n-1} + \xi_{n-1} \quad (18)$$

The unknown gain  $\alpha$  is re-estimated after each observation of  $x$ , the estimation being made according to the Bayesian formula,

$$\text{posterior density} = \text{likelihood} \times \text{prior density} \quad (19)$$

Since  $[\xi_n]$  is an independent Gaussian sequence, the successive posterior densities can be made Gaussian. The likelihood is given by

$$l(x_n|\alpha) = \exp \left[ -\frac{1}{2} \frac{(x_n - x_{n-1} - \alpha u_{n-1})^2}{\sigma} \right]$$

If the prior density is Gaussian with mean  $m_{n-1}$  and variance  $v_{n-1}$ , then eqn (19) gives

$$m_n = \frac{u_{n-1}(x_n - x_{n-1}) + m_{n-1} \frac{\sigma}{v_{n-1}}}{u_{n-1}^2 + \frac{\sigma}{v_{n-1}}}, \quad v_n = \frac{\sigma}{u_{n-1}^2 + \frac{\sigma}{v_{n-1}}} \quad (20)$$

Eqns (20) can now be used to up-date the mean and variance after each observation. It is to be noted that they are non-linear. In order to compute the control at instant  $n-1$ , it is necessary to have *a priori* distribution of the mean at the next time instant  $n$ . This can be found by substituting for  $x_n$  as yet unknown, in eqn (20) from (18) yielding a stochastic equation

$$m_n = \frac{m_{n-1} \frac{\sigma}{v_{n-1}} + \alpha u_{n-1}^2 + u_{n-1} \xi_{n-1}}{u_{n-1}^2 + \frac{\sigma}{v_{n-1}}} \quad (21)$$

The performance index with  $r$  stages to go (note  $r$  indexes time backwards) is taken as

$$f_r(x_r, m_r, v_r) = \min_u E \left\{ \sum_{k=r}^1 (u_k^2 + x_{k-1}^2) \right\} \quad (22)$$

Note the  $n$  in formulae (18) and (20) will also index backwards when used in conjunction with (22). The authors regret this notational inconsistency.

A discrete time iteration for the performance index may now be set up by using Bellman's Principle of Optimality

$$f_r(x_r, m_r, v_r) = \min_{u_r} E \{ x_{r-1}^2 + u_r^2 + f_{r-1}(x_{r-1}, m_{r-1}, v_{r-1}) \}$$

$$f_1(x_1, m_1, v_1) = \min_{u_1} E \{ x_0^2 + u_1^2 \} \quad (23)$$

It is understood that the value of  $f_{r-1}$  to be used in this iteration is the one resulting from the application of  $u_r$ .

On substituting for  $x_0$ , and averaging over both  $\xi$  and the current density of  $\alpha$  followed by minimization with respect to  $u_1$ ; the analytic expression for  $u_1$  and  $f_1$  may readily be found as

$$u_1 = \frac{-m_1 x_1}{m_1^2 + v_1 + 1} \quad f_1(x_1, m_1, v_1) = \frac{x_1^2(1 + v_1)}{m_1^2 + v_1 + 1} + \sigma \quad (24)$$

The expression for  $f_2$ , indexing  $n$  backwards, is now

$$f_2(x_2, m_2, v_2) = \min E \left\{ u_2^2 + (x_2 + \alpha u_2 + \xi_2)^2 \frac{A}{B} \right\} + \sigma \quad (25)$$

where

$$A = 2 + \frac{2\sigma}{u_2^2 + \frac{\sigma}{v_2}} + \frac{\left( m_2 \frac{\sigma}{v_2} + \alpha u_2^2 + u_2 \xi_2 \right)^2}{\left( u_2^2 + \frac{\sigma}{v_2} \right)^2}$$

$$B = 1 + \frac{\sigma}{u_2^2 + \frac{\sigma}{v_2}} + \frac{\left( m_2 \frac{\sigma}{v_2} + \alpha u_2^2 + u_2 \xi_2 \right)^2}{\left( u_2^2 + \frac{\sigma}{v_2} \right)^2}$$

It is evident that no simple analytic expression can be found for  $f_2$ . The complexity of the performance index expression is seen to be a consequence of the highly non-linear nature of the estimation equations; this is a common occurrence in adaptive systems. Now it is simple to evaluate  $f_2$  at any chosen point  $(x, m, v)$  using a digital computer, but this produces  $f_2$  as a set of numerical values; to avoid storing all these points  $f_2$  can be condensed into a set of three dimensional orthogonal polynomials. This was done on a small digital computer having 1,024 word working store using a programme developed by Cadwell and Williams<sup>5</sup>.

Cadwell and Williams' programme is designed for a particular small computer. It uses a modification of Forsyth's method for generating orthogonal polynomials of successively higher order using only the previous two polynomials. Owing to machine limitations only 200 data points can be fitted in three variables. However, in this example, it was found that the mean square error could be made less than 0.5 per cent when tested over a larger number of points. The polynomials were computed up to order 4, involving 35 coefficients of powers of  $x, m$ , and  $v$ .

Having approximated  $f_2$  it is possible to compute  $f_3$ . Since the error of approximation is small the iteration for  $f_3$  may be written

$$f_3(x_3, m_3, v_3) = \min_{v_3} E \{ u_3^2 + x_2^2 + f_2(x_2, m_2, v_2) + d_2(x_2, m_2, v_2) \} \quad (26)$$

where  $d_2(x_2, m_2, v_2)$  is the error in  $f_2$ , and  $f_3^*$  are the computed values of  $f_3$ . Now the minimum in  $f_3^*$  will be close to that in  $f_3$ , so that

$$f_3^* - f_3 + \min E \{ d_2 \} \quad (27)$$

Roughly, then, the error is additive at each stage.

To give an idea of the range of the coefficients of the powers of  $x, m, v$ , in  $f_3$  the largest value was 11.2, then there were nine in the range 1-10, eight in the range 0.1-1, eleven in the range 0.01-0.1 and only seven below this. Numerical experiments showed that omission of some of the smaller coefficients had a serious effect in certain regions of the variables.

The optimal value of control was found by a gradient technique, the expression for  $f$  being evaluated for a sequence of  $u$  values. A defect of the method is that polynomial approximations tend to show ripples, especially near the end of the fitted range, these ripples acted as false minima during the gradient computations, and it was necessary to check each minimum by approaching it from two sides. However, even this check was a little uncertain when the minimum was very flat, in practical terms this could make great differences to the control and was important. These effects were greatly reduced by approximating

the results so that the control was smooth function of  $x, m, v$ . Once  $f_3$  had been found at a suitable set of points a further polynomial approximation could be found and the whole process repeated to find  $f_4$ . The polynomial approximation method required the same computational process at each stage, which is convenient.

In this example after three stages the control became a stationary function of state, some results are shown in Figure 10. By trial and error a simple approximation for the stationary control was found to be

$$u \approx \frac{-mx}{m^2 + \frac{1}{2}v^2 + 1} \quad (28)$$

for  $x, m$  and  $v$  in the range 0–5.0.

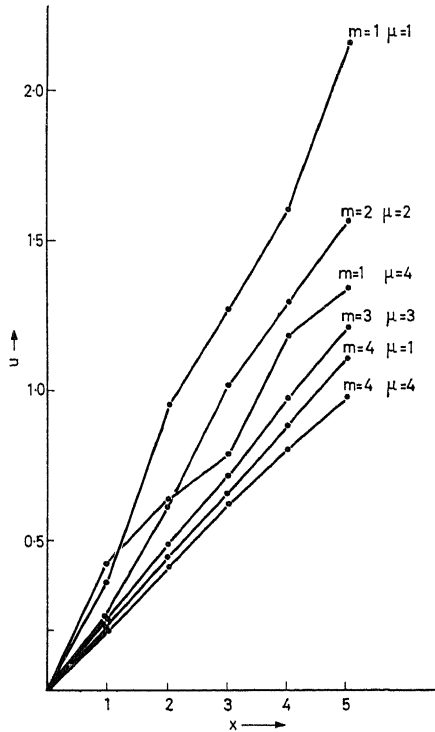


Figure 10. Stationary control law for adaptive regulator

On the particular computer used (300  $\mu$ sec multiplication and 4 msec division time) each point involved about 1 min of computation and over four stages some 800 points were required, needing in all some 14 h of computation, it is therefore interesting to see what further method of approximation could be used so as to reduce this computation load.

An alternative possibility is to replace the system by one which, on physical grounds, would appear to have a similar control solution. The simplest alternative system is to regard  $\alpha$  as a random variable with a fixed distribution at each stage, neglecting for the moment the transitions in mean and variance. Thus at each stage the only variable to be considered is  $x$ . The mean  $m$ , and the variance  $v$ , of the estimate of  $\alpha$ , are then successively up-dated and used in the solution for  $u$  (30).

The functional iteration for the stochastic system is a function of  $x$  only

$$f_r(x_r) = \min_{u_r} E \{x_{r-1}^2 + u_r^2 + f_{v-1}(x_{r-1})\} \quad (29)$$

$$f_1(x_1) = \min_{u_1} E \{x_0^2 + u_1^2\}$$

This iteration may be carried out analytically very simply. The stationary solution is

$$u = \frac{-cmv}{1+v+cv} \quad (30)$$

where  $c$  is the positive solution of

$$c = \frac{1}{2} \left( \frac{v}{m^2} - 1 \right) \pm \frac{1+v}{m} \quad (31)$$

On comparing the resulting values of control from eqn (28) and (30) it will be found the stochastic control is 10–20 per cent smaller than the adaptive control. This is to be expected on physical grounds, since the adaptive control has an exploratory element. However, the stochastic control value is a reasonable approximation to the adaptive one, considering the immense difference in the amount of computing involved.

At the present time the best approximation method for adaptive control known to the authors is the replacement of the fully adaptive system by the partial one as above. Since each unknown parameter often involves two, or more statistics, depending on the distribution used, the reduction in dimensionality can be substantial. It is usually apparent on physical grounds that the approximation will be a valid one. In many cases the accuracy of approximation will be better than in the example above.

#### A Function Space Approximation for Deterministic Systems

In non-linear deterministic systems a function space approximation can often be a useful alternative to the trajectory method. These methods depend on having an analytic form for the performance index of an approximated system. First it must be possible to rewrite the dynamic equation of the system in linearized form as

$$\dot{x} = A(x, t)x + B(x, t)u \quad (32)$$

To obtain an analytic form for the performance index choose

$$f(x, t) = \min_u \int_0^\infty \frac{1}{2} (x^T Q x + u^T R u) \cdot d\tau \quad (33)$$

where  $Q$  and  $R$  may be general time dependent matrices. However, for simplicity, take them to be constants in this description. The partial differential equation for  $f(x, t)$  is

$$-\frac{\partial f}{\partial t} = \min_u \left[ x^T \frac{Q}{2} x + u^T \frac{R}{2} u + p^T (A(x, t)x + B(x, t)u) \right] \quad (34)$$

On minimizing and regarding  $A$ , and  $B$  as constant it is found that

$$u = -R^{-1}B(x, t)^T p \quad (35)$$

where

$$p = [p_1, p_2, \dots, p_n]^T \quad \text{with} \quad p_i = \frac{\partial f}{\partial x_i}$$

After substituting for the optimal control it will be found that eqn (34) can be solved by substituting  $f(x, t) = \frac{1}{2} x^T P x$  where  $P$  is the solution of a matrix Riccati equation:

$$P' + A^T(x, t)P + PA(x, t) + Q = PB(x, t)R^{-1}B^T(x, t)P$$

Utilizing the stability properties of the Riccati equation<sup>7</sup> it can be shown that  $P$  is the positive definite solution of

$$PA(x, t) + A^T(x, t)P + Q = PB(x, t)R^{-1}B(x, t)^T P \quad (36)$$

The approximation scheme now uses eqn (36) to solve a state dependent matrix. The resulting matrix is substituted into eqn (35) to compute the control vector. The advantages of this method are (a) the solution requires only algebraic computation, and uses currently available quantities; (b) the precise nature of  $A$  and  $B$  is unimportant, thus the method is readily extendable to an adaptive scheme, where  $A$  and  $B$  vary as the result of measurement, and (c) it can be shown that an appropriate choice of the linearization will always result in a stable controlled system.

The stability and ease of realization of the resultant controlled system are probably the most important practical factors in favour of this technique. To show the stability, consider the general second order system

$$\begin{aligned} \dot{x}_1 &= a_{12}(x_2)x_2 \\ \dot{x}_2 &= a_{21}(x_1)x_1 + a_{22}(x_1, x_2)x_2 + du \end{aligned} \quad (37)$$

the factors  $a_{21}(x_1)$  and  $a_{22}(x_1, x_2)$  can be any bounded functions, but  $a_{12}(x_2)$  must satisfy certain requirements given below. Take the performance index

$$f(x, t) = \min_u \left[ \frac{1}{2} \int_t^T (x_1^2 + x_2^2 + u^2) d\tau \right] \quad (38)$$

Applying eqn (36) the elements  $p_{ij}$  of  $P$  satisfy

$$\begin{aligned} d^2 p_{11} p_{12} - a_{12} p_{11} - (a_{21} + a_{22}) p_{12} &= 0 \\ d^2 p_{12}^2 - 2 a_{21} p_{12} - 1 &= 0 \\ d^2 p_{22}^2 - 2 a_{22} p_{22} - 2 u_{12} p_{12} - 1 &= 0 \end{aligned} \quad (39)$$

On finding the stable solution of eqns (39) it is possible to evaluate the control

$$u = -d(p_{12}x_1 + p_{22}x_2)$$

where  $p_{12}$  and  $p_{22}$  are the positive solutions of eqns (39).

On substituting for  $u$  in the original equations the controlled system has the form:

$$\begin{aligned} \dot{x}_1 &= a_{12}(x_2)x_2 \\ \dot{x}_2 &= -b_{21}(x_1)x_1 - b_{22}(x_1, x_2)x_2 \end{aligned} \quad (40)$$

where the functions  $b_{21}$  and  $b_{22}$  are the positive roots of

$$\begin{aligned} b_{21}(x_1) &= [a_{21}^2(x_1) + d^2]^{\frac{1}{2}} \\ b_{22}(x_1, x_2) &= [a_{22}^2(x_1, x_2) + d^2(2a_{12}(x_2)p_{12} + 1)]^{\frac{1}{2}} \end{aligned} \quad (41)$$

The stability of eqns (40) can now be established by application of the second method of Liapunov<sup>8</sup>. Consider the proposed Liapunov function

$$V(x) = \int_0^{x_1} w b_{21}(w) dw + \int_0^{x_2} w a_{12}(w) dw \quad (42)$$

Clearly  $V(x)$  is a positive function which is bounded if  $a_{12}(x_2)$  is positive and suitably restricted. Its time derivative is

$$V' = -b_{22}(x_1, x_2)a_{12}(x_2)x_2^2 \quad (43)$$

which is negative semi-definite. If the  $x_1$  axis is not a permissible trajectory of the system of eqn (40) then the expression of eqn (42) is a Liapunov function of the system. This function defines a series of closed surfaces over the whole phase space about the origin which the trajectories enter. Thus the system is asymptotically stable about the origin.

A particular case which has been studied computationally is the Van der Pol equation

$$x'' + a(1 - x^2)x' + bx = du \quad (44)$$

This equation was linearized in phase space form by putting

$$\dot{x}_1 = x_2, \quad a_{12} = 1, \quad a_{21} = -b, \quad a_{22} = -a(1 - x_1^2) \quad (45)$$

Using the performance index of eqn (38) some comparisons of the computed trajectories in the exact and approximated cases are shown in Figure 11, which gives the phase space trajectories

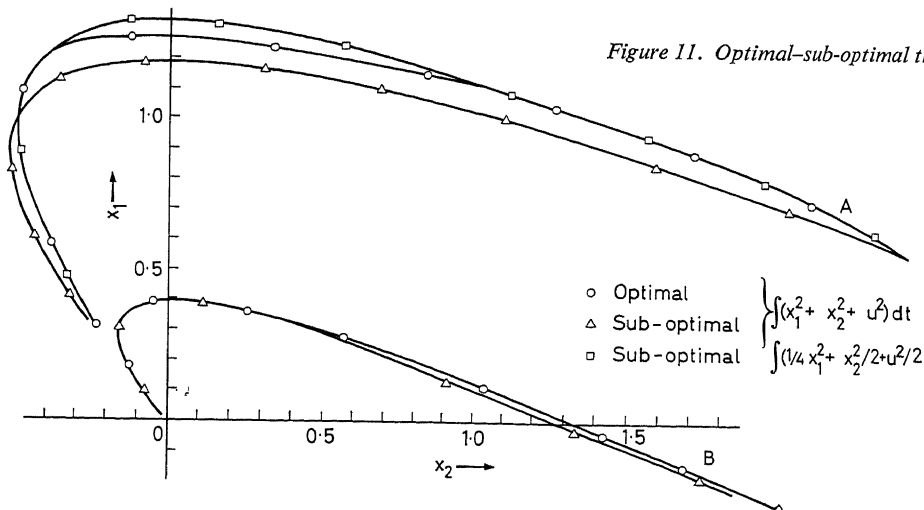


Figure 11. Optimal-sub-optimal trajectories



and the necessary control signals for the exact and approximated cases, Figure 12. The closeness of the approximation occurs in many practical cases, and is an indication of the effectiveness of the method. To illustrate the implementation of the scheme an analogue computer arrangement is shown in Figure 13, which solves eqn (39).

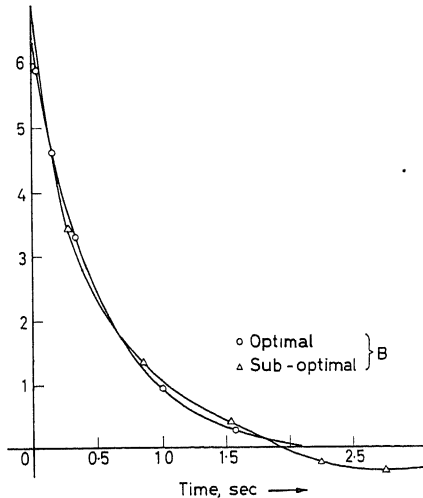


Figure 12. Optimal and sub-optimal control signals

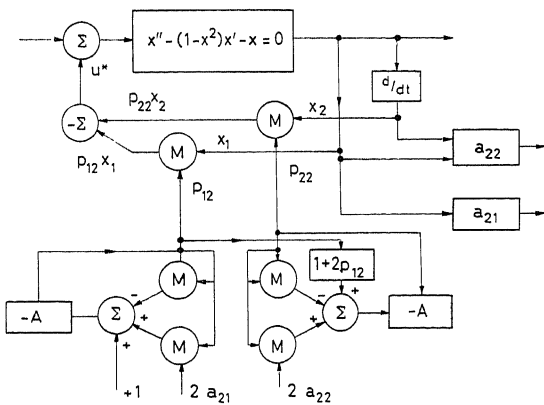


Figure 13. Analogue solution of sub-optimal system

The degree of approximation can be improved by varying the performance index slightly (i.e.  $Q$  and  $R$ ). Figure 4 shows the effects of such variations and Figure 11 suggests that a system linearized and optimized with respect to

$$f(x, t) = \int_t^\infty \left( \frac{1}{4} x_1^2 + \frac{1}{2} x_2^2 + \frac{1}{2} u^2 \right) dt$$

closely approximates the non-linear system optimized with respect to

$$f(x, t) = \frac{1}{2} \int_t^\infty (x_1^2 + x_2^2 + u^2) dt$$

Thus it appears possible to rescale the approximate phase plane to fit the optimal one by adjusting the performance index appropriately.

## Conclusions

The general mathematical equations for optimal and adaptive control can now be set up, but comprehensive methods for solution are not known. Approximation methods are being developed, but they must be used according to the individual circumstances of each problem. A number of different possible techniques of approximation have been given with appropriate examples.

The first point stressed is that the mathematical setting up of the problem can often be varied so as to make the computation easier, whilst still giving a satisfactory physical solution. The second point is that the methods can be grouped into two classes, trajectory methods applicable only to deterministic systems, and the function space method which is of wide application. A systematic search technique for solving trajectory problems has been described. A function space method using orthogonal expansions and another using linearized equations has also been given.

Computing was done at the University of London Computer Unit.

## Nomenclature

$\alpha$	Unknown gain factor
$A$	Term in dynamic equation
$B$	Term in dynamic equation
$d$	Control coefficient
$f(x, t)$	Performance index
$H$	Hamiltonian type function
$l(x, a)$	Likelihood function
$L$	Integrand in performance index
$m$	Estimated mean value
$n(p)$	Norm function
$p$	Co-state vector
$P$	Matrix in expansion of performance index
$Q$	Cost of state matrix
$R$	Cost of control matrix
$t, T$	Time
$u$	Control vector
$v$	Estimated variance
$V(x)$	Liapunov function
$x$	State vector
$\sigma$	Variance
$\{\xi_n\}$	Noise process
$\nu$	Lagrange multiplier

## References

- BELLMAN, R. *Adaptive Control*. 1961. Princeton; Princeton University Press
- PONTRYAGIN, L. S. Gamkrelidze Boltyanski and Mischenko. On the Theory of Optimal Processes. *C. R. Acad. Nauk SSSR* 110 (1956) 7
- FLORENTIN, J. J. Optimal control of continuous time Markov stochastic systems. *J. Electron. Control* 10, No. 6 (June 1961), 473
- ROZONER, L. I. L. S. Pontryagin's maximum principle. *Automation and Remote Control*. October, November, December 1959
- CADWELL, J. H., and WILLIAMS, D. E. Orthogonal methods of curve and surface fitting. *Computer Journal* 4, No. 3 (October 1961)
- LASALLE, J. P., and LEFSCHETZ, S. *Stability by Liapunov's Direct Method*. 1961. New York; Academic Press
- PEARSON, J. D. Approximation methods in optimal control. I. Sub-optimal control. *J. Electron. Control* 13, No. 5 (November 1962) 453
- WESTCOTT, J. H. *An Exposition of Adaptive Control*. 1962. London; Pergamon Press

## DISCUSSION

A. R. M. NOTON, *Electrical Engineering Department, University of Nottingham, Nottingham*

The authors have described the difficulties associated with the solution of the two-point boundary value problem, following from the application of Pontryagin's principle to deterministic systems. The trajectories are grossly sensitive to perturbations in the initial conditions and neither the digital nor the analogue computer provides a convenient method of solution.

In such computations the solutions are unrolled as functions of time but, to solve the two-point boundary value problem, another approach has been suggested<sup>1</sup>. The set of differential equations in  $x$  and  $p$  are approximated by their finite-difference form, i.e. the differential equations are replaced by a finite number of algebraic equations. These equations can then be solved on an analogue computer employing a corresponding number of operational amplifiers. In many practical cases, however, a significant saving on the required number of amplifiers can be achieved by preliminary algebraic manipulations and eliminations. The boundary values can easily be inserted and the method seems particularly attractive for approximating the required computations, at least on an on-line basis. Have the authors considered this approach?

## Reference

- <sup>1</sup> KIPINIAK, W. *Dynamic Optimization and Control*. 1961. New York; Wiley

J. H. WESTCOTT, J. J. FLORENTINE and J. D. PEARSON, *in reply*

The amount of analogue equipment necessary for the matrix technique appears to be excessive in general. However, this method may have special applications.

F. KYLSTRA, *Koninklijke/Shell Laboratorium, Badhuisweg 3, Amsterdam, Netherlands*

For the particular system described by eqns (9) the authors demonstrate the insensitivity of the solution to the formulation of the performance index. The fact is that this obviously holds in case transit time is the criterion. Is there some evidence that may lead one to generalize this observation to the point where, for a whole class of problems, one would be justified in employing a convenient index rather than the exact one, without trying both? It seems that this particular system already offers a counter example in the case of the control energy criterion.

J. H. WESTCOTT, J. J. FLORENTIN and J. D. PEARSON, *in reply*

Our point is that this is the case for some problems and is worth an investigation. Our criterion in this example was the 'similarity of trajectories' for various performance indices.

G. C. AGARWAL, *C.I.S.L. School of Electrical Engineering, Purdue University, Lafayette, Indiana, U.S.A.*

The authors have not given any reference to the quasi-linearization technique<sup>1</sup> for numerical solution of two-point boundary value problems. This technique is computationally very efficient with quadratic convergence property and can be applied to non-linear and also partial differential equations with multipoint boundary conditions. The method is particularly suitable for the numerical solution of the set of differential equations of the type in eqn (6). Many examples have been worked, using this method, at Control and Information Systems Laboratory, Purdue University, U.S.A.

The linearization technique eqns (32) and (36) which was proposed earlier (Reference 7 of the paper), was tried by me for a few first- and

second-order non-linear systems. The approximate solution was found to be very sensitive to the digital increment. No satisfactory result could be obtained.

## Reference

- <sup>1</sup> KALABA, R. On non-linear differential equations, the maximum operation and monotone convergence. *J. Math. Mech.* 8 (1959) 519

J. H. WESTCOTT, J. J. FLORENTIN and J. D. PEARSON, *in reply*

We have tried quasi-linearization techniques since this paper was written, and obtained some success. Digital computer solutions of non-linear differential equations are always 'sensitive to the digital increment'. The examples quoted gave us no difficulty.

H. NOUR-ELDIN, *Swiss Federal Institute of Technology, Gloriastrasse, Zurich 6, Switzerland*

I congratulate the authors on their interesting paper. It is really worth while to show the possibilities and limitations of the theory in solving concrete problems. In using the function space technique (dynamic programming), the authors overcame the shortage of the computer's storage by approximating the decision function by a three-dimensional orthogonal polynomial. They also showed that neglecting the transits in mean and variance will reduce to one variable which is the phase coordinate  $X$ . I would like to remark that the authors have a first-order system as an example. If the system is of higher order, one has to use Lagrange multipliers which are functions of time. The number of these multipliers will increase as the phase coordinates of the system increase and I feel that this is the real handicap of setting the problem for digital computers. Can the authors say whether any success regarding this point was achieved?

Furthermore, regarding optimal systems having performance indexes of the form of integral of sign definite functions ( $V$ ), I think some success can be achieved by constructing the Liapunov function corresponding to this sign definite function ( $V$ ). The solution of linear system leads to a system of algebraic equations. The authors have shown the advantages of this method. I mention that this method can be applicable to systems non-linear in phase coordinates but linear in the control variable. The differential equations can be written in the form of

$$\ddot{X} = F(X) + BU$$

One can find the Liapunov function for the unperturbed motion satisfying the equation

$$\sum_{i=1}^n \frac{\partial W}{\partial \dot{X}_i} \dot{X}_i = -V(x_1, \dots, x_n)$$

with  $V$  sign definite function and the performance index

$$I = \int_0^t V dt$$

This equation can be solved, using Zubov's method of construction of Liapunov functions, as long as the system differential equations fulfil Zubov's requirements.

J. H. WESTCOTT, J. J. FLORENTIN and J. D. PEARSON, *in reply*

We agree that the selection of Lagrange multipliers is a problem. The difficulty of a more exact analysis using Zubov's method, for example, is that the complexity of the controller increases rapidly with the degree of Zubov's polynomial expansion.

V. W. EVELEIGH, *General Electric Co., 3-118 Electronics Park, Syracuse, N.Y., U.S.A.*

The authors have presented an interesting summary of techniques available for solving optimization and adaptive control problems. The examples illustrating how a solution may often be obtained by proper choice of the performance criterion are particularly well chosen.

The memory requirements often encountered in the application of dynamic programming techniques, and the possibility of using a polynomial approximation to overcome this difficulty, as indicated by the authors, has also been pointed out by Aoki<sup>1</sup>, Kipiniak<sup>2</sup>, Peterson<sup>3</sup>, and Merriam<sup>4</sup>, among others. Unfortunately, as also pointed out by the authors, very little has been accomplished in developing a theory of error for these approximate techniques. Work in this direction may well prove fruitful in the future.

In the area of computational techniques, I would like to indicate an entirely different line of development than that referred to in the paper. It is currently possible, using any one of several alternate techniques available, to obtain the optimum solution to a very general class of non-linear and/or time-varying differential equations with boundary values at two or more points. The method involved an iterative procedure based upon linearization of the equations about a nominal (non-optimum) solution. A form of this method was perhaps first conceived by Breakwell<sup>5</sup>, but Kipiniak<sup>2</sup> also suggests a similar technique. Breakwell's original suggestion was to proceed to the optimum solution in one iteration, but it is often found that this approach fails due to the range of extrapolation required. Kelley<sup>6</sup> and, independently, Bryson and Denham<sup>7</sup>, have developed iterative techniques based upon linearization of the solution about the nominal path, but in which the size of the improvement is constrained to assure efficient convergence to the desired solution. More recently, Merriam<sup>4</sup>, and Breakwell, Speyer and Bryson<sup>8</sup>, have developed second variation techniques with improved convergence capabilities near the desired optimum. Also, techniques for dealing with inequality constraints upon one or more elements of the state vector have been developed by Dreyfus<sup>9</sup>, and by Bryson and Denham<sup>10</sup>. An excellent set of examples illustrating the use of these techniques is available in the literature. Each of the methods, with minor additional development or interpretation, gives rise to a control law designed to force the system to follow an optimum or near optimum trajectory through phase space. Analytical control examples are also available, but few, if any, actual hardware applications have been reported.

I hasten to point out that this short bibliography of computational techniques is not offered in criticism of the author's survey of tech-

niques available, but only as a supplement. I am fully aware of the long lead time required between the work upon which a paper of this type is based and its presentation at the conference. The authors are no doubt already aware of some or all of these references and have found, or will find them most interesting.

#### References

- <sup>1</sup> AOKI, M. *Dynamic Programming and Numerical Experimentation as Applied to Control*, Nov. 1959 PhD. Thesis, U.C.L.A.
- <sup>2</sup> KIPINIAK, W. *Dynamic Optimization and Control*. 1961. London and New York; M.I.T. Press and John Wiley
- <sup>3</sup> PETERSON, E. L. *Statistical Analysis and Optimization of Systems*. 1961. London and New York; John Wiley
- <sup>4</sup> MERRIAM, C. W. III. *Optimization Theory and the Design of Feedback Control Systems*. 1964. New York; McGraw-Hill
- <sup>5</sup> BREAKWELL, J. V. The optimization of trajectories, *J. Soc. Industr. Appl. Math.* 7 (1959)
- <sup>6</sup> KELLEY, H. J. Gradient theory of optimal flight paths. *J. Amer. Rocket Soc.* 30 (1960)
- <sup>7</sup> BRYSON, A. E. and DENHAM, W. F. A steepest ascent method for solving optimum programming problems. *J. Appl. Math. Mech., Leningr.* (June 1962)
- <sup>8</sup> BREAKWELL, J. V., SPEYER, J. L. and BRYSON, A. E. Optimization and control of non-linear systems using the second variation, *J. Control Sec. Ind. Appl. Math.* Ser. A, 1, No. 2 (1963)
- <sup>9</sup> DREYFUS, S. *Variational Problems With State Variable Inequality Constraints*. Rand Corp. Paper No. P-2605 (July 1962)
- <sup>10</sup> BRYSON, A. E. and DENHAM, W. F. *The Solution of Optimal Programming Problems with Inequality Constraints*. Nov. 1962. Raytheon Document Number BR-2121, Missile and Space Div., Bedford, Mass.

J. H. WESTCOTT, J. J. FLORENTIN and J. D. PEARSON, *in reply*

Dr. Eveleigh questions the stability of the approximation procedure of replacing an adaptive system by a stochastic one. In use, the uncertainty, i.e. the variance of the estimated parameters, would be artificially increased. This would lead to a control less dependent on the numerical values of the parameter estimate, thus counteracting instability.

We thank him for his additional survey of current computational methods.

# An Optimal Guidance Approximation for Quasi-circular Orbital Rendezvous

H. J. KELLEY and J. C. DUNN

## Summary

Three-dimensional guidance about a time-optimal rendezvous flight path is examined within the framework of a quasi-circular orbital dynamics assumption. The guidance scheme, optimal in the same sense as the nominal trajectory, is based formally upon the approximate construction of a field of neighbouring optimal rendezvous paths, following a method developed by Kelley in an earlier paper<sup>1</sup>.

For purposes of illustration, a class of reference trajectories, generated by a low magnitude, circumferentially directed thrust vector, has been adopted for the guidance analysis. Such trajectories are time optimal, under appropriate circumstances, for transfer between neighbouring co-planar circular orbits. The application considered permits an analytical representation of the extremal field which in turn leads to a closed form linear feedback control solution with time-varying gains.

Some suggestions are also given for possible modifications which might enhance system accuracy and the range of operability during practical implementation of low-thrust rendezvous guidance.

## Sommaire

On examine le problème du guidage dans les trois dimensions au voisinage d'une trajectoire de rendez-vous optimale par rapport au temps, dans le cadre de la dynamique des orbites quasi-circulaires. Le schéma de guidage, qui est optimal dans le même sens que la trajectoire nominale, est formellement basé sur la construction approximative d'un champ de trajectoires de rendez-vous optimales voisines, suivant la méthode exposée par M. Kelley dans un article récent.

Comme exemple, on a adopté pour l'analyse du guidage une classe de trajectoires de référence produite par un vecteur circonferential de faible poussée. De telles trajectoires sont, dans des circonstances appropriées, optimales par rapport au temps pour le transfert entre des orbites circulaires voisines dans le même plan. L'application considérée autorise une représentation analytique du champ extrémal qui, à son tour, mène à une solution de commande sous forme linéaire fermée à réaction, avec gains variant dans le temps.

En conclusion, on présente quelques suggestions, qui ont, pour objet d'améliorer la précision du système et le domaine de fonctionnement dans la réalisation pratique du guidage de rendez-vous à faible poussée.

## Zusammenfassung

Das Problem der 3-dimensionalen Lenkung, um eine zeitoptimale Flugbahn für Begegnungen von Raumfahrzeugen (Rendezvous-Technik) zu erzielen, wird im Rahmen einer quasi-kreisförmig angenommenen (dynamischen) Bahn untersucht. Der Lenkvorgang, im gleichen Sinne optimal wie die Bezugsbahn, stützt sich formal auf die angenäherte Konstruktion eines Bereiches (Bündel) von nebeneinanderliegenden optimalen Begegnungsbahnen; diese Methode wurde in einem früheren Aufsatz von Kelley entwickelt.

Zur Erläuterung wurden eine Reihe von Begegnungsbahnen, die ein in Richtung der Peripherie zeigender Schubvektor von geringem Betrag erzeugt, für die Untersuchung der Lenkung angenommen. Solche Flugbahnen sind zeitoptimal, sie eignen sich unter entsprechenden Bedingungen zum Überwechseln zwischen zwei nebeneinanderliegenden, in der gleichen Ebene befindlichen kreisförmigen Bahnen.

Die betrachtete Anwendung läßt eine analytische Darstellung des extremen Bereiches zu, was auf einen Regelkreis mit zeitveränderlicher Verstärkung führt.

Der Aufsatz enthält einige Vorschläge für mögliche Abänderungen, die die Genauigkeit und den Operationsbereich während der praktischen Lenkung mit geringem Schub erhöhen können.

## Introduction

The second-order guidance approximation scheme employed in this paper has been developed in an earlier publication<sup>1</sup>. Essentially, the idea is to select a flight path, optimized in some appropriate sense, as a nominal trajectory, and then to base guidance upon a family, or 'field', of optimal trajectories approximated in the vicinity of the nominal.

The investigation, as is shown later, applies this scheme to the guidance problem for orbital rendezvous. The method is tractable for a fairly wide class of problems, although in general it must be carried through numerically. However, in the present case an analytical treatment becomes feasible because of the recent availability of a particularly simple optimal transfer manoeuvre suitable for use as a nominal trajectory. The nominal manoeuvre, as it appears in this paper, is a direct outgrowth of a co-planar circular orbit transfer analysis conducted by Hinz<sup>2</sup>. The problem posed by Hinz, although phrased somewhat differently with regard to coordinate systems and assumptions employed in deriving the equations of motion, is mathematically equivalent to the nominal transfer manoeuvre problem investigated herein. Both cases yield to an analytical treatment of the boundary value problem whenever the manoeuvre duration is an integral multiple of a reference orbit's period.

The analysis of three-dimensional rendezvous guidance for the class of trajectories discussed above leads directly to a synthesis of a linear feedback control solution with time-varying gains given in closed form.

Some suggestions are also given for possible modifications which might enhance system accuracy and the range of operability during practical implementation of low-thrust rendezvous guidance.

## The Differential Equations of Powered Flight

Considerations begin with the differential equations of three-dimensional powered flight in a central inverse-square force field (Figure 1):

$$\begin{aligned} \ddot{r} - r \cos^2 \psi \dot{\theta}^2 - r \dot{\psi}^2 + \frac{k}{r^2} &= \frac{F}{m} \cdot v \cos \beta \sin \alpha \\ \ddot{\theta} + 2 \frac{\dot{r} \dot{\theta}}{r} - 2 \tan \psi \dot{\psi} \dot{\theta} &= \frac{F}{mr \cos \psi} \cdot v \cos \beta \cos \alpha \end{aligned} \quad (1)$$

$$\ddot{\psi} + 2\frac{\dot{r}\dot{\psi}}{r} + \frac{1}{2}\sin 2\psi \dot{\theta}^2 = \frac{F}{mr} \cdot v \sin \beta$$

$$\dot{m} = -\frac{vF}{C}; \quad -\frac{\pi}{2} \leq \beta \leq \frac{\pi}{2}, \quad 0 \leq v \leq 1$$

where  $F$  is the maximum thrust level of the reaction engine;  $C$  is the propellant exhaust velocity;  $v$  is a throttle variable; and  $m$  is the instantaneous vehicle mass. The difficulty in obtaining any sort of particular solution for these equations needs no comment here, except that it provides a motive for the simplifications which are now introduced. The object is to devise certain assumptions which will allow the replacement of eqns (1) by an approximate set of differential equations which are linear in the state variables  $r$ ,  $\theta$ , and  $\psi$ , and their time derivatives and, preferably, separable in the control variables  $v$ ,  $\alpha$ , and  $\beta$ . (Simplifications of this kind are required to make flight path optimization and guidance problems analytically tractable.) To be specific, it is preferable that these approximate differential equations describe low-thrust acceleration transfer trajectories between neighbouring circular orbits.

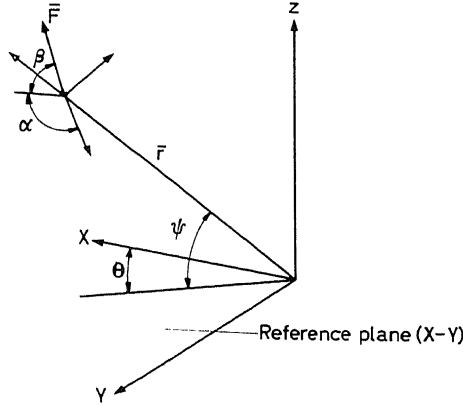


Figure 1. Coordinate geometry

The following set of dependent and independent variable transformations will prove useful for our purposes. Let

$$r(t) = R_0 [1 + \eta(t)]$$

$$\theta(t) = \tau(t) + \varepsilon(t)$$

and

$$m(t) = m_0 [1 + \xi(t)]$$

where  $R_0$  is the radius of a circular reference orbit situated in the  $\psi = 0$  plane (Figure 1);  $m_0$  is some reference mass; and  $\tau$  is a fictitious angle defined by the differential expression

$$\frac{d\tau}{dt} = \omega_0, \quad \tau(0) = 0$$

where  $\omega_0$  is the reference orbit's period.

Furthermore, since  $\tau(t)$ , as defined above, is a monotonic time-like parameter, it is permissible to change the independent variable in eqns (1) from  $t$  to  $\tau$ . This can be accomplished by simply relating  $t$  derivatives to  $\tau$  derivatives as follows:

$$\frac{d}{dt} = \frac{d\tau}{dt} \frac{d}{d\tau} = \omega_0 \frac{d}{d\tau}$$

$$\frac{d^2}{dt^2} = \frac{d}{d\tau} \left( \frac{d}{d\tau} \right) = \omega_0^2 \frac{d^2}{d\tau^2}$$

Finally, if note is made of the fact that  $k/R_0^3 \omega_0^2 = 1$ , then eqns (1) can be put into the following reduced first-order form:

$$u' = (1 + \eta)(1 + v)^2 \cos^2 \psi - \frac{1}{(1 + \eta)^2} + \frac{F^*}{(1 + \xi)} v \cos \beta \sin \alpha$$

$$v' = -\frac{2u(1 + v)}{1 + \eta} + 2w(1 + v) \tan \psi$$

$$+ \frac{F^*}{(1 + \eta)(1 + \xi)} v \cos \beta \cos \alpha$$

$$w' = -\frac{2uw}{1 + \eta} - \frac{1}{2}(1 + v)^2 \sin 2\psi + \frac{F^*}{(1 + \eta)(1 + \xi)} v \sin \beta$$

$$\eta' = u \quad (2)$$

$$\varepsilon' = v$$

$$\psi' = w$$

$$\xi' = -\frac{F^*}{C^*} v$$

where  $F^* = F/m_0 R_0 w_0^2$  is the reduced maximum thrust acceleration;  $C^* = C/R_0 w_0$  is the reduced exhaust velocity; and the superscribed prime denotes differentiation with respect to the reduced time,  $\tau$ .

Now let us assume that  $F^*$ ,  $\xi$ ,  $u$ ,  $v$ ,  $w$ ,  $\eta$ , and  $\psi$  are all terms of order  $\delta$  or smaller ( $\delta < 1$ ). Under these circumstances, one would therefore anticipate that all terms of order  $\delta^2$  in eqns (2) will become negligible with respect to terms of order  $\delta$ . Thus the following simplified differential equations are arrived at:

$$u' = 2v + 3\eta + F^* v \cos \beta \sin \alpha$$

$$v' = -2u + F^* v \cos \beta \cos \alpha$$

$$w' = -\psi + F^* v \sin \beta \quad (3)$$

$$\eta' = u$$

$$\varepsilon' = v$$

$$\psi' = w$$

It may be said that, to the first order of small quantities, eqns (3)\* are descriptive of quasi-circular flight for the following reason; if, as has been assumed,  $u$ ,  $v$ ,  $\eta$ , etc. are of order  $\delta$ , then it follows at once that

$$\left[ \left( \frac{E - E_0}{E_0} \right)^2 + \left( \frac{h - h_0}{h_0} \right)^2 \right]^{\frac{1}{2}} = 0(\delta)$$

where  $E$  and  $h$  are specific energy and angular momentum respectively. Consequently, the energy-momentum images of

\* These equations are identical in form to the differential equations of Wheelon<sup>3</sup> and Anthony<sup>4</sup>. However, the dependent variables and thrust vector steering angles are not subject to the same interpretation. In particular, the quantity  $\varepsilon$  in eqns (3) is not required to be small—an important point in the subsequent development.

trajectories which are adequately described by eqns (3) should everywhere be close to the locus of circular orbits in the  $E$ - $h$  phase plane (Figure 2).

Clearly, the validity of the quasi-circular differential equations will be compromised when the parameters  $F^*$ ,  $F^*/C^*$ , and  $\tau$  exceed certain critical values. Just precisely what these critical values are cannot be determined until the nature of the control schedule  $v(\tau)$ ,  $\alpha(\tau)$ , and  $\beta(\tau)$ , is specified. The reader is advised to bear this in mind in the sequel.

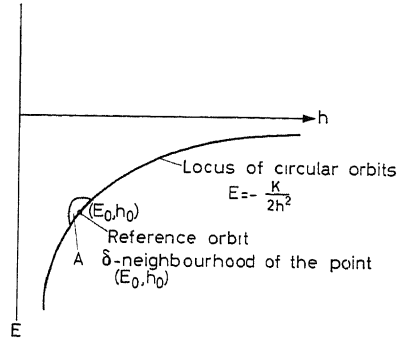


Figure 2. Energy-momentum phase diagram

### Optimal Transfer Between Neighbouring Circular Orbits

The optimal orbit transfer problem may be stated as follows: given two neighbouring circular orbits, find the steering angles  $\alpha(\tau)$  and  $\beta(\tau)$  and the throttle schedule  $v(\tau)$  which produce a transfer between the two orbits in minimum time. For present purposes, the case for which the terminal orbits are co-planar will be selected, a class of optimal transfer paths within the framework of the quasi-circular dynamics assumption derived, and later, these paths employed as nominal trajectories for the three-dimensional rendezvous guidance analysis.

To reiterate, if the subscripts 0 and  $f$  denote initial and final conditions respectively, a search is made for a set of control functions  $\alpha(\tau)$ ,  $\beta(\tau)$ , and  $v(\tau)$  which minimize  $\tau_f$ , produce a state transition which evolves in accord with eqns (3), and which satisfies the circular orbit boundary conditions, viz: at  $\tau = 0$ ,  $u = v = w = \eta = \varepsilon = \psi = 0$ ,  $\xi = 0$ ; at  $\tau = \tau_f$ ,  $u = w = 2v + 3\eta = \psi = 0$ ,  $\eta = K(\text{const.})$ . The problem so stated is a Mayer variational problem with bounded control variables. The necessary conditions to be satisfied by its solution are well known and are written here for the present application without further comment:

Let

$x_i$  = state variables,  $u, v, w, \eta, \varepsilon, \psi, \xi$

$y_k$  = control variables  $v, \alpha, \beta$

$\lambda_i$  = multiplier functions

$A_i$  = undetermined constant multipliers

$P$  = function to be extremized =  $\tau_f + A_1(u_f) + A_2$

$$(2v_f + 3\eta_f) + A_3(w_f) + A_4(\eta_f - K) + A_5(\psi) \quad (4)$$

$H$  = Hamiltonian function =  $\sum_i \lambda_i x_i'$

Then the following equations and inequalities must be satisfied,

$$H(y) \geq H(\bar{y}) \quad (5)$$

i.e., the optimal control  $\bar{y}$  minimizes the function  $H$ .

$$\lambda_i' = -\frac{\partial H}{\partial x_i} \quad (6)$$

$$x_i' = \frac{\partial H}{\partial \lambda_i} \quad [\text{eqns (3)}] \quad (7)$$

together with the corresponding natural boundary and transversality conditions,

$$H_f = -\frac{\partial P}{\partial \tau_f} \quad (5a)$$

$$\lambda_{i_f} = \frac{\partial P}{\partial x_{i_f}} \quad (6a)$$

$$x_{i_0} = 0, \quad i = 1, \dots, 7$$

$$x_{1_f} = x_{3_f} = 2x_{2_f} + 3x_{4_f} = x_{6_f} = 0, \quad x_{4_f} = K \quad (7a)$$

Now, a minimum of  $H$  is attained at a minimum of  $H_1$ , where  $H_1$  is that part of  $H$  which depends on the control variables,  $y$ . For the problem here,

$$H_1 = F^* v \left( \lambda_1 \cos \beta \sin \alpha + \lambda_2 \cos \beta \cos \alpha + \lambda_3 \sin \beta - \frac{\lambda_7}{C^*} \right) \quad (8)$$

However, note that the final mass,  $m_0(1 + \xi_f)$  does not appear in the pay-off, i.e., the final mass is left open. However, only those trajectories for which  $|\xi_f|$  is of order  $\delta$  are admissible because of assumptions implicit in eqns (3). Therefore, eqns (6) and (6a) imply that  $\lambda_7 \equiv 0$ . Consequently, eqn (8) simplifies to:

$$H_1 = F^* v (\lambda_1 \cos \beta \sin \alpha + \lambda_2 \cos \beta \cos \alpha + \lambda_3 \sin \beta) \quad (8a)$$

The requirements on the control variables  $\alpha$ ,  $\beta$ , and  $v$  are determined by reasoning as follows.

Since  $\alpha$  is unbounded,  $\partial H_1 / \partial \alpha = 0$  and  $\partial^2 H_1 / \partial \alpha^2 \geq 0$  at the minimum of  $H_1$  and hence,

$$\sin \bar{\alpha} = -\lambda_1 / (\lambda_1^2 + \lambda_2^2)^{1/2}, \quad \cos \bar{\alpha} = -\lambda_2 / (\lambda_1^2 + \lambda_2^2)^{1/2} \quad (9)$$

$H_1$  reduces to:

$$H_1 = F^* v [-(\lambda_1^2 + \lambda_2^2)^{1/2} \cos \beta + \lambda_3 \sin \beta] \quad (10)$$

which can be written in the form

$$H_1 = F^* v (\lambda_1^2 + \lambda_2^2 + \lambda_3^2)^{1/2} \sin(\beta - \varphi) \quad (11)$$

$$\varphi = \sin^{-1} \frac{(\lambda_1^2 + \lambda_2^2)^{1/2}}{(\lambda_1^2 + \lambda_2^2 + \lambda_3^2)^{1/2}}$$

By virtue of the fact that  $\sin \varphi \geq 0$ , it follows that the principal value of  $\varphi$  lies between 0 and  $\pi$ . This last conclusion, together with eqn (11), permits the deduction that the minimum of  $H_1$  occurs when  $\beta = \varphi - \pi/2^*$ , i.e., when

$$\sin \bar{\beta} = -\frac{\lambda_3}{(\lambda_1^2 + \lambda_2^2 + \lambda_3^2)^{1/2}}, \quad \cos \bar{\beta} = \frac{(\lambda_1^2 + \lambda_2^2)^{1/2}}{(\lambda_1^2 + \lambda_2^2 + \lambda_3^2)^{1/2}} \quad (12)$$

$$\left( -\frac{\pi}{2} \leq \beta \leq \frac{\pi}{2} \right)$$

and  $H_1$  reduces still further to:

$$H_1 = -F^* v (\lambda_1^2 + \lambda_2^2 + \lambda_3^2)^{1/2} \quad (13)$$

\* Notice that  $H_1$  is also stationary with respect to this value of  $\beta$ , i.e.,  $(\partial H_1 / \partial \beta) \bar{\beta} = 0$ .

From eqn (13) it follows immediately that  $\nu = 1$  minimizes  $H_1$  whenever

$$(\lambda_1^2 + \lambda_2^2 + \lambda_3^2)^{\frac{1}{2}} \neq 0 \quad (14)$$

Furthermore, it can be verified [by solving eqns (6)] that  $\lambda_1^2 + \lambda_2^2 + \lambda_3^2 \neq 0$  except at a finite number of points on any  $\tau$  interval of length  $2\pi$ . Thus the indeterminate values of  $\nu$  corresponding to  $\lambda_1^2 + \lambda_2^2 + \lambda_3^2 = 0$  form a set of measure zero and our problem is therefore well behaved.

In summary, the optimal control variables depend upon the multipliers  $\lambda_1, \lambda_2$ , and  $\lambda_3$  in the following manner:

$$\begin{aligned} \sin \bar{\alpha} &= -\frac{\lambda_1}{(\lambda_1^2 + \lambda_2^2)^{\frac{1}{2}}}, & \cos \bar{\alpha} &= -\frac{\lambda_2}{(\lambda_1^2 + \lambda_2^2)^{\frac{1}{2}}} \\ \sin \bar{\beta} &= -\frac{\lambda_3}{(\lambda_1^2 + \lambda_2^2 + \lambda_3^2)^{\frac{1}{2}}}, & \cos \bar{\beta} &= \frac{(\lambda_1^2 + \lambda_2^2)^{\frac{1}{2}}}{(\lambda_1^2 + \lambda_2^2 + \lambda_3^2)^{\frac{1}{2}}} \end{aligned} \quad (15)$$

$\nu = 1$

The  $\lambda$ 's in turn depend on the unknown Lagrange multipliers,  $\lambda_i$  through eqn (6), and the solution of the boundary value problem turns upon one's capability to solve for these undetermined constants. However, before attention is directed to the boundary value problem, it is worth while to emphasize two points. First, it is easily demonstrated that the optimal control variables given by eqns (15) satisfy a strong form of eqn (5), i.e.,

$$H_1(y) > H(\bar{y}), \quad y \neq \bar{y} \quad (16)$$

But eqn (16) together with the linear character of eqns (3), are sufficient conditions for a strong relative minimum of  $P$ . Thus it is assured that the control law of eqns (15) will provide a time-optimal transfer between two co-planar circular orbits which is unique whenever the boundary value equations possess a unique solution. Second, it should be noted that the result expressed by the last of eqns (15) is independent of the multiplier functions  $\lambda_i$ , and is therefore insensitive to the boundary conditions. Thus a full throttle operational mode is a characteristic of the entire extremal field of quasi-circular transfer trajectories for the problem of minimum time transfer with final mass open.

#### The Orbital Transfer Boundary Value Problem—a Special Class of Solutions for Rendezvous

The differential equations, eqns (3), when written in a symbolic matrix notation, have the following structure:

$$x' = Ax + g(y) \quad (17)$$

where  $A$  is a matrix of constant coefficients and  $g$  is a vector whose elements depend upon the control variables  $\alpha, \beta$ , and  $\nu$ . Solutions for eqn (17) can therefore usually be phrased in terms of a fundamental solution matrix  $\Phi(\tau_f, \tau)$  and superposition integrals, e.g.:

$$x(\tau_f) = \Phi(\tau_f, 0)x(0) + \int_0^{\tau_f} \Phi(\tau_f, \tau)g[y(\tau)]d\tau \quad (18)$$

Eqn (19) \*

$$g = \{g_1, \dots, g_7\}^T$$

where

$$\begin{aligned} g_1 &= F^* \nu \cos \beta \sin \alpha = -\frac{\lambda_1 F^*}{(\lambda_1^2 + \lambda_2^2 + \lambda_3^2)^{\frac{1}{2}}} \\ g_2 &= F^* \nu \cos \beta \cos \alpha = -\frac{\lambda_2 F^*}{(\lambda_1^2 + \lambda_2^2 + \lambda_3^2)^{\frac{1}{2}}} \\ g_3 &= F^* \nu \sin \beta = -\frac{\lambda_3 F^*}{(\lambda_1^2 + \lambda_2^2 + \lambda_3^2)^{\frac{1}{2}}} \\ g_4 &= g_5 = g_6 = 0 \\ g_7 &= \frac{F^* \nu}{C^*} = -\frac{F^*}{C^*} \end{aligned} \quad (20)$$

As already pointed out, the  $\lambda$ 's depend on the undetermined constants  $\lambda_i$  in accord with eqn (6a). Furthermore, it is noted that in matrix notation, eqns (6) have the form

$$\lambda' = -A^T \lambda \quad (21)$$

and thus are adjoint differential expressions for eqn (17) [i.e., eqns (3)]. Consequently, their solution is determined when  $\Phi(\tau_f, \tau)$  is known, i.e.,

$$\lambda(\tau) = \Phi^T \lambda(\tau_f) \equiv \Phi^T \left\{ \frac{\partial P}{\partial x_{if}} \right\} \quad (22)$$

where

$$\frac{\partial P}{\partial x_{1f}} = \lambda_1, \quad \frac{\partial P}{\partial x_{2f}} = 2\lambda_2 \text{ etc.}$$

In view of these considerations, the co-planar circle-to-circle transfer boundary conditions [see eqn (7)] are:

\* Eqn (19):

$$\Phi(\tau_f, \tau) = \begin{bmatrix} \cos \hat{\tau} & 2 \sin \hat{\tau} & 0 & 3 \sin \hat{\tau} & 0 & 0 & 0 \\ -2 \sin \hat{\tau} & -(3 - 4 \cos \hat{\tau}) & 0 & -6(1 - \cos \hat{\tau}) & 0 & 0 & 0 \\ 0 & 0 & \cos \hat{\tau} & 0 & 0 & -\sin \hat{\tau} & 0 \\ \sin \hat{\tau} & 2(1 - \cos \hat{\tau}) & 0 & (4 - 3 \cos \hat{\tau}) & 0 & 0 & 0 \\ -2(1 - \cos \hat{\tau}) & -(3 \hat{\tau} - 4 \sin \hat{\tau}) & 0 & -6(\hat{\tau} - \sin \hat{\tau}) & 1 & 0 & 0 \\ 0 & 0 & \sin \hat{\tau} & 0 & 0 & \cos \hat{\tau} & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 \end{bmatrix} \quad (19)$$

where: (a)  $\tau_0$  is the point in time at which corrective action is initiated; (b) the  $\delta x_{i_f}$ 's are dispersions which would occur at the terminus of the nominal trajectory in the absence of corrective guidance; (c)  $\delta t_f$  is a variation in the final time,  $\bar{\tau}_f$  ( $\delta\tau_f \ll 1$ ); and (d) the coefficients of  $\delta\tau_f$  in each equation are to be interpreted as state variable closing rates, i.e., differences between vehicle and target state variable derivatives on the nominal path at  $\tau = \bar{\tau}_f$  (notice that only  $v$  has a non-zero closing rate).

A total of six equations linear in the six unknowns  $C_1, \dots, C_5$ , and  $\delta\tau_f$  are at our disposal. Their solution may be obtained most conveniently by first combining the second and fourth of eqns (34) to obtain the result,

$$\delta v_f + 2\delta\eta_f = -F^* \delta\tau_f \quad (35)$$

which may then be used to eliminate  $\delta\tau_f$ . The remaining equations can then be written in matrix notation (see Appendix), e.g.:

$$A(\tau_0) \begin{Bmatrix} C_1 \\ C_2 \\ C_3 \end{Bmatrix} = J \begin{Bmatrix} \delta u \\ \delta v \\ \delta\eta \\ \delta\epsilon \end{Bmatrix}_f \quad (36)$$

and

$$B(\tau_0) \begin{Bmatrix} C_4 \\ C_5 \end{Bmatrix} = K \begin{Bmatrix} \delta w \\ \delta\psi \end{Bmatrix}_f \quad (37)$$

where  $A(\tau_0)$  and  $B(\tau_0)$  are matrices whose elements depend upon  $\tau_0$  and  $J$  and  $K$  are constant matrices. But the  $\delta x_{i_f}$  are related to the  $\delta x_{i(\tau_0)}$  through eqn (19). Therefore the constants  $C_i$  may be expressed as functions of the 'initial' excursions at time  $\tau_0$ , i.e.:

$$\begin{Bmatrix} C_1 \\ C_2 \\ C_3 \end{Bmatrix} = A^{-1}(\tau_0) J \Phi_1(\tau_0) \begin{Bmatrix} \delta u \\ \delta v \\ \delta\eta \\ \delta\epsilon \end{Bmatrix}_{\tau_0} \quad \det |A| \neq 0 \quad (36a)$$

$$\begin{Bmatrix} C_4 \\ C_5 \end{Bmatrix} = B^{-1}(\tau_0) K \Phi_2(\tau_0) \begin{Bmatrix} \delta w \\ \delta\psi \end{Bmatrix}_{\tau_0} \quad \det |B| \neq 0 \quad (37a)$$

where  $\Phi_1$  and  $\Phi_2$  are submatrix elements of the complete state transition matrix [eqn (19)]. Eqns (32), (36a), and (37a) then provide the open loop solution for the steering angle corrections, viz:

$$\delta\alpha_\tau = \{\cos \hat{\tau}, \sin \hat{\tau}, 1\} A^{-1}(\tau_0) J \Phi_1(\tau_0) \begin{Bmatrix} \delta u \\ \delta v \\ \delta\eta \\ \delta\epsilon \end{Bmatrix}_{\tau_0} \quad (38)$$

$\tau_0 \leq \tau \leq \bar{\tau}_f$

and

$$\delta\beta_\tau = \{\cos \hat{\tau}, \sin \hat{\tau}\} B^{-1}(\tau_0) K \Phi_2(\tau_0) \begin{Bmatrix} \delta w \\ \delta\psi \end{Bmatrix}_{\tau_0} \quad (39)$$

$\tau_0 < \tau < \bar{\tau}_f$

In order to obtain the closed loop linear feedback control solution mentioned earlier, it is only necessary to replace  $\tau_0$  by  $\tau$  in eqns (38) and (39), which are then applicable for  $0 \leq \tau \leq \bar{\tau}_f$ .

### Considerations on Rendezvous Guidance Implementation

In this section, several ways to enhance system accuracy and range of operability in practical implementation of low-thrust rendezvous guidance are suggested.

First of all it is noted that real time generation of the system gains and the state variables along the nominal trajectory will be required in practice. In the case of command guidance, these would presumably be calculated 'on the fly' by a ground-based computer, whereas if a vehicle-borne computer were employed, these quantities would more likely be stored in polynomial approximation. Now, with regard to the computation of gains, it is worth calling attention to the advisability, in either case, of employing double precision arithmetic in the calculation of the determinants of the Appendix, since they tend rapidly toward zero as time-to-go approaches zero. This, of course, does not necessarily imply the need for a highly precise representation of the gains themselves.

As concerns the generation of the state variables along the nominal trajectory, it should be pointed out that over the course of many revolutions there will be an accumulation of error stemming from higher-order effects if the linear equations are used. This error will place an additional burden upon the guidance system. The situation may be alleviated somewhat by introducing second-order corrections to the nominal in the following approximate fashion. Second-order terms of the first two equations of the system (5), which are significant along the nominal, are

$$\begin{aligned} p &= -3\eta^2 + 2\eta v + v^2 \\ q &= -2uv + 2u\eta - F^*\eta \end{aligned} \quad (40)$$

If  $p(t)$  and  $q(t)$  are estimated by insertion of the nominal values of  $u, v$ , and  $\eta$  as computed via the linearized equations, corrections  $\delta u_e, \delta v_e, \delta\eta_e$  and  $\delta\epsilon_e$ , corresponding to the forcing terms  $p$  and  $q$ , may then be calculated by means of the influence functions presented earlier. The results take a particularly simple form if the integrals are evaluated at  $\tau = 2m\pi$ ,  $m$  an integer,  $0 \leq m \leq n$ , namely

$$\begin{aligned} \delta u_i(2m\pi) &= 22F^{*2}m\pi \\ \delta v_i(2m\pi) &= F^{*2}(68m^2\pi^2 - 80m\pi) \\ \delta\eta_i(2m\pi) &= F^{*2}(-16m^2\pi^2 + 40m\pi) \\ \delta\epsilon_i(2m\pi) &= F^{*2}(36\pi + 8m^3\pi^3) \end{aligned} \quad (41)$$

If these are employed as corrections to the nominal values with  $2m\pi$  replaced by  $\tau$ , the result is incorrect by the omission of certain oscillatory effects; however, it is thought that these will be unimportant in comparison with the *secular* terms which are properly accounted for. In any event, the idea is to provide the closed loop control system with some anticipation for errors accruing on account of non-linearity.

One of the most significant limitations arising from the approximations made in the foregoing analysis is the restriction that the change in arrival time from the nominal be small compared with the orbital period. Shift in arrival time is directly proportional to the difference between actual and nominal values of the linearized energy parameter,  $v + 2\eta$ . Furthermore, analysis reveals that relatively large steering corrections are associated with errors in this variable. Since the energy parameter is a monotonically increasing function of time, it therefore appears reasonable to consider this



parameter as a candidate for independent variable in the mechanization of the system. According to this scheme, the system gains and the state variables along the nominal would be generated (or stored) as functions of energy-to-be-gained, the difference between the terminal and instantaneous values of  $v + 2\eta$ . In an analysis conducted with an independent variable having fixed terminal value, complexities arising from shifts in terminal time would be avoided and there would be one less approximation required.

There is another point to which scant attention has been paid in the present investigation, which is essentially a feasibility study, and this is the choice of reference orbit and the position of the target vehicle in relation to the reference axis system. We have tacitly assumed the target to be moving in a perfectly circular orbit and chosen our axis system location for analytical convenience as being at the initial point of the nominal orbit transfer trajectory which terminates at the target. It seems clear for a number of reasons that the appropriate choice in actual implementation of a system would be a reference circular orbit having the period of the target's orbit and a reference axis system moving along this circular orbit in the vicinity of the target. One advantage of such a choice is the facilitation of guidance in terms of relative position and velocity measurements.

Finally, it is pointed out that by regarding  $F^*$  as an auxiliary state variable (as opposed to a parameter) defined by the differential equation,  $F^{*'} = 0$ , it then becomes possible, with little real

increase in complexity, to derive an additional term for the guidance law whose function would be to counteract the effects of instantaneous fluctuations in the reduced thrust-acceleration level. The procedure for treating system parameters as auxiliary state variables is discussed at greater length by Kelley<sup>1</sup>.

## Conclusions

The low-thrust quasi-circular orbital rendezvous example, presented in this paper, illustrates a possible practical application of an optimal guidance scheme developed in a previous publication. The closed-form result obtained for guidance corrections should be safely applicable for correctional manoeuvres during the last several revolutions of a low-thrust ascending (or descending) approach to rendezvous. Some numerical computations designed to establish the range of validity of the guidance approximation are currently in progress.

*The authors extend their gratitude to Mr. Frank Sobierajski and Mrs. Agnes Zevens of Grumman's Research Department: to the former, for assistance in preparing and verifying several of the foregoing analytical results; to the latter, for the preparation of the figures appearing in this paper.*

*This research was partially supported by the USAF Office of Scientific Research under Contracts AF 29(600)—2671 and AF 49(638)—1207.*

## Appendix

The matrices appearing in eqns (36) through (39) are:

$$J = \frac{1}{F^*} \begin{bmatrix} -1 & 0 & 0 & 0 \\ 0 & 0 & -1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix}$$

$$K = -\frac{1}{F^*} \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$$

$$A = \begin{bmatrix} \left[ \frac{(2n\pi - \tau_0)}{2} - \frac{(\sin 2\tau_0)}{4} \right] & \frac{(1 - \cos 2\tau_0)}{4} & -\sin \tau_0 \\ \frac{(1 - \cos 2\tau_0)}{4} & \left[ \frac{(2n\pi - \tau_0)}{2} + \frac{(\sin 2\tau_0)}{4} \right] & (1 - \cos \tau_0) \\ \left[ \frac{(\sin 2\tau_0)}{2} - [2n\pi - \tau_0] - 2\sin \tau_0 \right] & (1 - \cos \tau_0)^2 & (4n\pi - 2\tau_0 + 2\sin \tau_0) \end{bmatrix}$$

$$B = \begin{bmatrix} \left( \frac{2n\pi - \tau_0}{2} - \frac{\sin 2\tau_0}{4} \right) & \left( \frac{1 - \cos 2\tau_0}{4} \right) \\ \left( \frac{1 - \cos 2\tau_0}{4} \right) & \left( \frac{2n\pi - \tau_0}{2} + \frac{\sin 2\tau_0}{4} \right) \end{bmatrix}$$

$$\Phi_1 = \begin{bmatrix} \cos \tau_0 & -2\sin \tau_0 & 3\sin \tau_0 & 0 \\ 2\sin \tau_0 & -(3 - 4\cos \tau_0) & -6(1 - \cos \tau_0) & 0 \\ -\sin \tau_0 & 2(1 - \cos \tau_0) & (4 - 3\cos \tau_0) & 0 \\ -2(1 - \cos \tau_0) - (6n\pi - 3\tau_0 + 4\sin \tau_0) - (12n\pi - 6\tau_0 + \sin \tau_0) & & & 1 \end{bmatrix}$$

$$\Phi_2 = \begin{bmatrix} \cos \tau_0 & \sin \tau_0 \\ -\sin \tau_0 & \cos \tau_0 \end{bmatrix}$$

also if

$$x = \frac{2n\pi - \tau_0}{2}, 0 \leq x \leq n\pi$$

$$\det |A| = 4x^3 - 4\sin^2 x (3 - \sin^2 x)x + 8\sin^3 x \cos x \neq 0$$

for  $x > 0$

$$\det |B| = x^2 - \frac{1}{4}(\sin^2 2x) \neq 0 \text{ for } x > 0$$

## References

- <sup>1</sup> KELLEY, H. J. Guidance theory and extremal fields, *Inst. Radio Engrs. Nat. Aerospace Electronics Conf.*, Dayton, Ohio (May 1962); *Trans. Inst. Radio Engrs., AC* (October 1962)
- <sup>2</sup> HINZ, H. K. Optimal low-thrust near-circular orbital transfer, *AIAA Jnl.* (in the press)
- <sup>3</sup> WHEELON, A. D. Midcourse and terminal guidance, *Space Technology*. 1959. New York; Wiley
- <sup>4</sup> CLOHESSY, W. H. and WILTSHIRE, R. S. Terminal guidance system for satellite rendezvous, *J. Aerospace Sci.* Vol. 27, No. 9 (September 1960)
- <sup>5</sup> BREAKWELL, J. V. and BRYSON, A. E. Neighboring-optimum terminal control for multivariable nonlinear systems, *Meeting Soc. Industr. Appl. Math.*, Cambridge, Massachusetts (November 1962)

# On Synthesizing Optimal Controls

L. W. NEUSTADT and B. H. PAIEWONSKY

## Summary

A synthesis procedure for a wide class of optimal processes is presented. Linear processes are considered under the following types of optimization criteria and constraints: (1) Under control amplitude constraint, either minimize control effort or minimize time. (2) Under effort constraint, minimize time. (3) Under both amplitude and effort constraints, minimize time. (4) Under no constraint, minimize effort. In each case, the control system is to take the state vector from a given initial state to a given terminal state, or to a given terminal closed convex set. Control effort is understood either in the sense of maximum amplitude or of an integral of a certain function of the control.

It is well known that such optimal controls can be given in terms of a solution of the adjoint system of differential equations of the original system. The main difficulty in computing the optimal control lies in finding the initial conditions of the adjoint system. By geometric arguments, a function of the initial condition is developed, and it is shown that this function attains a maximum at the desired value of the initial condition. Furthermore, the gradient of this function is shown to have a simple form, so that the maximum can be directly computed by the method of steepest ascent.

## Sommaire

Une procédure de synthèse pour une large classe de processus optimaux est présentée. Des processus linéaires sont envisagés avec les types suivants de critères d'optimisation et de limitations: 1) Avec la limitation de l'amplitude de commande, soit la minimalisation de l'effort de commande, soit la minimalisation du temps. 2) Avec la limitation de l'effort, minimalisation du temps. 3) Avec la limitation aussi bien de l'amplitude que de l'effort, minimalisation du temps. 4) Sans aucune limitation, minimalisation de l'effort. Dans chaque cas le système de commande doit amener le vecteur d'état à partir d'un état initial donné jusqu'à un état final donné ou bien jusqu'à un ensemble convexe fermé final donné. L'effort de commande est compris dans le sens soit de l'amplitude maximum soit de l'intégrale d'une certaine fonction de la commande.

Il est bien connu que de telles commandes optimales peuvent être données sous forme d'une solution du système adjoint d'équations différentielles du système original. La principale difficulté de calcul de la commande optimale réside dans la détermination des conditions initiales du système adjoint. À l'aide d'arguments géométriques une fonction des conditions initiales est établie et il est montré que cette fonction atteint un maximum pour la valeur désirée des conditions initiales. Il est montré, ensuite, que le gradient de cette fonction possède une forme simple, de sorte que le maximum peut être directement calculé par la méthode de la pente la plus rapide.

## Zusammenfassung

Ein Syntheseverfahren wird erläutert, welches für eine große Klasse optimaler Prozesse gültig ist. Lineare Prozesse werden nach folgenden Gesichtspunkten der Optimierungskriterien und Beschränkungen untersucht: 1. Bei Begrenzung der Amplitude der Stellgröße mache man entweder die Regelleistung oder die Zeit zu einem Minimum; 2. bei Begrenzung in der Leistung suche man das Minimum der Zeit; 3. liegen Begrenzungen in der Amplitude und in der Leistung vor, mache man die Zeit zum Minimum; 4. sind keinerlei Begrenzungen vorhanden, suche man das Minimum der Leistung.

In jedem Fall hat das System den Zustandsvektor von einem gegebenen Anfangszustand bis zu einem bestimmten Endzustand zu durchlaufen, oder bis zu einer bestimmten geschlossenen, konvexen Endfläche. Die Regelleistung wird im Sinne eines Scheitelwertes oder eines Integrales über eine Funktion der Regelgröße verstanden.

Bekanntlich können derartige optimale Regelungen als Lösung des adjungierten Systems der Differentialgleichungen des ursprünglichen Systems angeschrieben werden. Die Hauptschwierigkeit bei der Berechnung der optimalen Regelung besteht im Auffinden der Anfangsbedingungen des adjungierten Systems. Hier wird nach geometrischen Überlegungen eine Funktion der Anfangsbedingung entwickelt; man kann zeigen, daß diese Funktion bei dem gewünschten Wert der Anfangsbedingung ein Maximum erreicht. Ferner ergibt sich, daß der Gradient dieser Funktion eine einfache Gestalt hat, so daß das Maximum unmittelbar nach der Methode des „steilsten Anstieges“ berechenbar ist.

## Introduction

The structure of optimal control processes is the subject of a great deal of current research. The maximum principle first formulated by Pontryagin, Boltyanskii, and Gamkrelidze<sup>1</sup> has been one of the fundamental results. This principle makes it possible to find optimal control functions for a wide class of optimal processes in terms of a solution of a linear differential equation which is related to the differential equation which describes the given control system. A difficulty arises, however, if one wishes actually to compute the optimal control, because the maximum principle does not prescribe the initial conditions necessary to solve the auxiliary differential equation. This paper presents an iterative procedure for computing these initial values for certain particular problems.

The method which is set forth is applicable to systems which can be described by a linear system of ordinary differential equations of the form:

$$\dot{x} = A(t)x + B(t)u(t) \quad (1)$$

where  $x$  is the state vector (say of dimension  $n$ ) in the phase space of the system,  $u(t)$  is the control vector (say of dimension  $r$ ), and  $A$  and  $B$  are  $n \times n$  and  $n \times r$  matrices which are continuous in the time  $t$ .

Suppose that a 'control region'  $U$  in  $r$  space is given, and that the control functions  $u(t)$  are constrained to take on their values in  $U$ . Vector functions with range in  $U$ , each of whose components are measurable, will be called admissible controls. If  $x(t)$  is a solution of eqn (1) with  $x(0) = x_0$  and  $u(t) = u^*(t)$ , and if  $x(t') = x_1$  for some  $t' > 0$ , it will be said that  $u^*(t)$  transfers  $x_0$  to  $x_1$  in time  $t'$ .

Finally, suppose that a scalar valued function  $\phi(u)$ , whose domain contains  $U$ , is given.  $\phi$  will be called the 'effort function', and if  $u(t)$  is an admissible control function, the functional

$$\int_0^T \phi(u(t)) dt \equiv \varepsilon(u(t))$$

( $T$  represents the duration of the control process) will be referred to as the control effort associated with  $u(t)$ . Physically,  $\varepsilon$  may represent fuel or gas consumption, electrical power loss, etc.

Given two points  $x_0$  and  $x_1$  in phase space, we shall consider the following optimization problems: (1) find an admissible control which transfers  $x_0$  to  $x_1$  in minimum time; (2) find an admissible control subject to the constraint  $\varepsilon(u(t)) \leq M$  (where  $M$  is a given constant) which transfers  $x_0$  to  $x_1$  in minimum time; (3) given a time  $T > 0$ , find an admissible control which transfers  $x_0$  to  $x_1$  in time  $T$  and which in so doing minimizes the effort  $\varepsilon(u(t))$ . In each case we shall have to assume that  $A, B, U$ , and  $\phi$  satisfy certain conditions. These, however, include a number of important physical problems.

According to the maximum principle, the optimal control for each problem is given in terms of a solution of the equation

$$\dot{\psi} = -A^* \psi \quad (2)$$

(where  $A^*$  is the adjoint matrix of  $A$ ). To find the proper initial conditions for eqn (2), we shall define a real valued function  $F(\eta) = F(\eta^1, \dots, \eta^n)$  such that  $F$  attains its maximum at those values of  $\eta$  for which the solution of eqn (2), with  $\psi(0) = -\eta$ , determines the optimal control. We shall also give an iterative procedure for computing this maximizing value. The method of computation is particularly suitable for hybrid analogue digital computers.

The synthesis procedure described in this paper exploits one of the prime features of the maximum principle: the optimal control is determined by  $n$  (or  $n+1$ ) real numbers—the initial values of the adjoint system. Our results are an extension of a procedure previously developed for the so-called ‘bang-bang’, time-optimal problem<sup>2</sup>.

## I. Time-Optimal Controls with no Effort Constraint

The time-optimal control problem with no effort constraint will, in general, have a solution only if  $U$  is compact. It will be assumed that  $U$  is compact and convex, and contains the origin as an interior point. First consider the case with  $x_1 = 0$ . This is physically the most important case.

The general solution of eqn (1) with initial value  $x(0) = x_0$  is given by

$$x(t) = X(t) \left[ x_0 + \int_0^t X^{-1}(s) B(s) u(s) ds \right] \quad (3)$$

where  $X(t)$  is the  $n \times n$  matrix function satisfying the equations

$$\dot{X}(t) = A(t) X(t); \quad X(0) = I \text{ (the identity)}$$

Let

$$C(t) = \left\{ - \int_0^t X^{-1}(s) B(s) u(s) ds : u(s) \text{ admissible} \right\} \quad (4)$$

It is clear that  $C(t)$  consists of those points which can be transferred to 0 in time  $t$ , using an admissible control. Because of the hypotheses on  $U$ , it follows immediately that  $C(t)$  is bounded, convex, contains the origin, and that  $C(t') \supset C(t)$  for  $t' > t$ . It can be shown that  $C(t)$  is closed (a proof for the special case of  $U$  a polyhedron is given by Pontryagin *et al.*<sup>1</sup>, and this proof is easily extended to  $U$  any compact, convex set). Finally, it follows from the compactness of  $U$  that if  $x \in C(t)$  for all  $t > t'$ , then  $x \in C(t')$ . It then immediately follows that there is a smallest  $t^*$  for which  $x_0 \in C(t^*)$ ; i.e., there is a time-optimal

control which transfers  $x_0$  to 0 in the minimum time  $t^*$ . Furthermore,  $x_0$  is a boundary point of  $C(t^*)$ . (See Reference 4, Lemma 3. It is easily seen that the hypotheses of this lemma are satisfied in the present case.) Therefore, there is a support plane to  $C(t^*)$  at  $x_0$ . Let  $\eta_* = (\eta_*^1, \dots, \eta_*^n)$  be the vector normal to this plane directed away from  $C(t^*)$ , so that

$$\eta_* \cdot x_0 \geq \eta_* \cdot x \quad \text{all } x \in C(t^*)$$

i. e., if

$$x_0 = - \int_0^{t^*} X^{-1}(t) B(t) u^*(t) dt$$

then

$$\int_0^{t^*} -\eta_* \cdot X^{-1}(t) B(t) u^*(t) dt \geq \int_0^{t^*} -\eta_* \cdot X^{-1}(t) B(t) u(t) dt \quad (5)$$

for every admissible  $u(t)$ . Note that the vector function  $-\eta_* \cdot X^{-1}(t)$  satisfies eqn (2) with initial value  $\psi(0) = -\eta_*$ . (The geometric arguments of this section were first given by Bellman, *et al.*<sup>3</sup> for a particular case of this problem, and have since been used by various other authors.)

One further assumption is now made concerning  $A, B$ , and  $U$ . It is supposed that if  $\psi(t)$  is any non-trivial solution of eqn (2), the function  $[\psi(t) B(t)] \cdot v$  of the vector variable  $v$  has a unique maximum in  $U$  for almost every  $t$ ,  $0 \leq t < \infty$ . (Since the origin is an interior point of  $U$ , this maximum is positive.) This condition is referred to here as the unique maximum condition. If  $U$  is the cube given by  $|u^j| \leq 1$ ,  $j = 1, \dots, n$  [where  $u = (u^1, \dots, u^n)$ ], this condition corresponds to the concept of normal control systems introduced by La Salle<sup>4</sup>. If  $U$  is an arbitrary, convex, closed polyhedron, this condition is equivalent to the general position condition described by Pontryagin *et al.*<sup>1</sup>.

The following notations are now introduced: Let  $\psi(t, \eta)$  be the solution of eqn (2) with initial value  $\psi(0) = -\eta$ . Let  $u(t, \eta)$  be the admissible control function which has the property that

$$[\psi(t, \eta) B(t)] \cdot u(t, \eta) = \max_{v \in U} [\psi(t, \eta) B(t)] \cdot v \quad (6)$$

By hypothesis,  $u(t, \eta)$  is uniquely defined for almost all  $t > 0$ . (We shall disregard sets of measure zero when describing control functions.)

It follows from eqn (5) that  $u^*(t) = u(t, \eta_*)$ . (This statement is a special case of the Pontryagin maximum principle.) Thus, if  $\eta_*$  is known, the solution  $\psi(t, \eta_*)$  can be computed from eqn (2), as a result of which the optimal control  $u^*(t)$  is uniquely determined by the maximum principle.

A method for finding  $\eta_*$  is now given. Let

$$z(t, \eta) = - \int_0^t X^{-1}(s) B(s) u(s, \eta) ds \quad (7)$$

for any non-zero vector  $\eta$  and any  $t > 0$ ;  $z(t, \eta) \in C(t)$ . By definition of  $u(t, \eta)$ , and by the unique maximum condition, it follows that

$$\eta \cdot z(t, \eta) > \eta \cdot \zeta \quad \text{for all } \zeta \in C(t), \quad \zeta \neq z(t, \eta) \quad (8)$$

Consider the function

$$f(t, \eta; x_0) = \eta \cdot [z(t, \eta) - x_0] \quad (9)$$

For fixed  $\eta \neq 0$ ,  $f$  is a continuous, strictly monotonically increasing function of  $t$  (since the maximum given by eqn (6) is positive). Let the domain of  $f$  be restricted to those  $\eta$  for which  $\eta \cdot x_0 = -f(0, \eta; x_0) > 0$ . If  $z(t^*, \eta) \neq x_0$ ,  $f(t^*, \eta; x_0) > 0$  by eqn (8), and, consequently,  $f(t, \eta; x_0) = 0$  for some unique  $t \in (0, t^*)$ . On the other hand, if  $z(t^*, \eta) = x_0$ , then  $f(t^*, \eta; x_0) = 0$  and  $f(t, \eta; x_0) < 0$  for  $t < t^*$ . Let  $F(\eta; x_0)$  be the value of  $t$  which satisfies the equation

$$f(F(\eta; x_0), \eta; x_0) = 0 \quad (10)$$

Then  $F$  is defined, and  $0 < F(\eta; x_0) \leq t^*$ . Furthermore,  $F(\eta; x_0) = t^*$  if, and only if,  $z(t^*, \eta) = x_0$  in which case [see eqns (3) and (7)]  $u(t, \eta)$  is the time-optimal control which transfers  $x_0$  to 0 in time  $t^*$ . Note that the vector  $\eta$  which determines the optimal control is not unique. Nevertheless, the optimal control is unique almost everywhere.

In summary the following has been shown: given a control system described by eqn (1), a convex, compact control region  $U$  which contains 0 as an interior point, and an initial point in phase space  $x_0$ . If there is an admissible control which transfers  $x_0$  to 0, and if the unique maximum condition is satisfied, then there is a unique time-optimal control  $u(t, \eta)$  defined by eqn (6), where  $\psi(t, \eta)$  is the solution of eqn (2) with  $\psi(0) = -\eta$  and  $\eta$  is any vector for which the function  $F$  defined by eqns (7), (9), and (10) takes on its maximum value  $t^*$ .

Now suppose that  $x_1$  is not the origin. In fact it can be supposed that the 'target point' is a (continuous) function of time  $\xi(t)$ . In addition, it can be supposed that the right-hand side of eqn (1) has an additional (continuous) 'forcing term'  $h(t)$ , and remove the restriction that  $0 \in U$ . Then, it can still be shown that if there is an admissible control which transfers  $x_0$  to  $\xi(t)$  in some time  $t$ , then there is a time-optimal control of the form  $u(t, \eta)$ , where  $\eta$  has an analogous geometric interpretation. An analogue of the function  $F$  can then be constructed; but in general,  $F$  will only be defined in a neighbourhood of the optimal values of  $\eta$ . However,  $F$  will again have a maximum at these (and only these) optimal values of  $\eta$ . For a more detailed discussion see Neustadt<sup>2</sup>.

## II. Time-Optimal Controls with Effort Constraint

Now consider the time-optimal problem in the presence of effort constraint. The discussion will be confined to the case where the 'target'  $x_1$  is the origin; the extension to the general case can be carried out as indicated above.

In this section it need not be assumed that  $U$  is bounded (in fact, it may be the entire  $r$  dimensional space, in which case there is no constraint on the allowed values of the control). It will be supposed, however, that  $U$  is closed and convex and contains the origin. The effort function  $\phi(u)$  will be assumed to be continuous, bounded from below on  $U$ , and to satisfy the relation  $\phi(0) = 0$ . Further implicit restrictions on  $\phi$  and  $U$  are stated below.

Corresponding to the set  $C(t)$  of Section I, the set  $C(t)$  in  $(n+1)$ -space will be defined as follows

$$C(t) = \left\{ \left( -\int_0^t X^{-1}(s) B(s) u(s) ds, \int_0^t \phi(u(s)) ds \right) : u(s) \text{ admissible} \right\}$$

Clearly,  $C(t)$  consists of the points  $(x^1, \dots, x^n, x^{n+1})$  such that the point  $(x^1, \dots, x^n)$  can be transferred to the origin in time  $t$  (using an admissible control) with control effort  $x^{n+1}$ . Note that  $C(t') \supset C(t)$  if  $t' > t$ .

Two further assumptions concerning  $A$ ,  $B$ ,  $\phi$ , and  $U$  will be made. First it is assumed that  $C(t)$  (for every  $t \geq 0$ ) is closed and convex and has the property that if  $x \in C(\tau)$  for all  $\tau > t$ , then  $x \in C(t)$ ; and secondly, it is assumed that a generalized unique maximum condition is satisfied: for every non-trivial solution  $\psi(t)$  of eqn (2), and every non-positive constant  $\beta$ , the function

$$[\psi(t) \cdot B(t)] \cdot v + \beta \phi(v)$$

of the vector variable  $v$  has a unique maximum in  $U$  for almost every  $t$ ,  $0 \leq t < \infty$ . (Since  $\phi(0) = 0$ , and  $0 \in U$ , this maximum is non-negative.)

Three important cases for which the first assumption is valid are the following:

(1) If  $U$  is the unit cube  $|u^j| \leq 1$ ,  $j = 1, \dots, r$ , and  $\phi(u)$  is a non-negative convex function which is defined (and continuous) in an open set which contains  $U$ , and which takes on its maximum value in  $U$  at each vertex of  $U$ ,  $C(t)$  satisfies the above hypotheses<sup>5</sup>. In particular, functions  $\phi(u)$  of the form  $\sum_{j=1}^r \lambda_j |u^j|^p$ , where the  $\lambda_j$  and  $(p-1)$  are non-negative constants, fall under this category.

(2) If  $U$  is the unit sphere  $\sum_{j=1}^r (u^j)^2 = \|u\|^2 \leq 1$ , and  $\phi(u) = \|u\|$ , the set  $C(t)$  has the desired properties.

(3) If  $U$  is the entire space, and  $\phi(u) = \sum_{j=1}^r \lambda_j |u^j|^p$  where the  $\lambda_j$  and  $(p-1)$  are positive, the first assumption is also valid.

The generalized unique maximum condition is more difficult to verify in particular examples. In general, this condition will be violated only if the time-optimal control is not unique. A more detailed discussion is given by Neustadt<sup>5</sup>.

Corresponding to the function  $u(t, \eta)$ , the function  $\hat{u}(t, \eta)$  will be defined, where  $\eta$  is any  $(n+1)$ -vector  $(\eta^1, \dots, \eta^n, \eta^{n+1}) = (\eta, \eta^{n+1})$  with  $\eta \neq 0$  and  $\eta^{n+1} \leq 0$ . Namely, let  $\hat{u}$  be the admissible control function which has the property that

$$[\psi(t, \eta) B(t)] \cdot \hat{u}(t, \eta) + \eta^{n+1} \phi(\hat{u}(t, \eta)) = \max_{v \in U} [\psi(t, \eta) B(t) v + \eta^{n+1} \phi(v)] \quad (11)$$

Here  $\psi(t, \eta)$  is defined as in Section I, and, by hypothesis,  $\hat{u}(t, \eta)$  is uniquely defined for almost all  $t > 0$ .

Now let  $l$  be the line segment in  $(n+1)$ -space defined by

$$l = \{x = (x_0, \xi) : \xi \leq M\}$$

where  $x_0$  is the given initial position in phase space, and  $M$  is the maximum allowed value of the control effort. Clearly, there is an admissible control which transfers  $x_0$  to 0 in time  $t$  with control effort not exceeding  $M$  if, and only if,  $C(t)$  meets  $l$ . Thus, one is looking for the smallest value of  $t$  for which  $C(t)$  meets  $l$ . Because of the assumptions on  $C(t)$  and  $\phi(u)$ , it follows that there is such a smallest  $t$ ; i.e., there is a time-optimal control satisfying the effort constraint. Denote the minimum time by  $t^*$ . It can also be shown that because each  $C(t)$  is convex, every point of  $C(t^*) \cap l$  is on the boundary of  $C(t^*)$ . The generalized unique maximum condition implies that every boundary point of  $C(t^*)$  is an extremal point of the set. Hence, there is only one point  $x^* = (x_0, \xi) \in l \cap C(t^*)$ . Now there are two possible cases: either  $\xi = M$  or  $\xi < M$ . If  $\xi < M$ ,  $C(t^*)$  meets the line

through  $(x_0, 0)$ , parallel with the  $x^{n+1}$  axis, only at  $x^*$ , and  $C(t)$ , for  $t < t^*$ , does not meet this line at all. Thus,  $t^*$  is the minimal transfer time from  $x_0$  to 0 in the absence of the effort constraint. The problem then reduces to that considered in Section I. Henceforth it is assumed that  $\xi = M$ , so that  $x^* = (x_0, M)$ .

Clearly,  $x^*$  is on the boundary of  $C(t^*)$ , and there is a support plane to  $C(t^*)$  at  $x^*$ . Let the normal to this plane that is directed away from  $C(t^*)$ , be  $\eta_* = (\eta_*^1, \dots, \eta_*^n, \beta) = (\eta_*, \beta)$ . It is clear from the geometry that  $\beta \leq 0$ . As in Section I, it follows that

$$x^* = (x_0, M) = \left( - \int_0^{t^*} X^{-1}(t) B(t) \hat{u}(t, \eta_*) dt, \int_0^{t^*} \phi(\hat{u}(t, \eta_*)) dt \right)$$

and  $\hat{u}(t, \eta_*)$  is the desired optimal control. Since this control is determined by the solution  $\psi(t, \eta_*)$  and  $\beta$  [see eqn (11)], and since these quantities are in turn determined by  $\eta_*$ , the optimal control can be computed once  $\eta_*$  is known.

As in Section I, a function is defined which takes on its maximum at the desired values of  $\eta_*$ .

Let

$$z(t, \eta) = \left( - \int_0^t X^{-1}(s) B(s) \hat{u}(s, \eta) ds, \int_0^t \phi(\hat{u}(s, \eta)) ds \right)$$

for every vector  $\eta = (\eta, \eta^{n+1})$  for which  $\eta \neq 0$  and  $\eta^{n+1} \leq 0$ . Since  $\hat{u}$  is admissible,  $z(t, \eta) \in C(t)$ . By virtue of the generalized unique maximum condition

$$\eta \cdot z(t, \eta) > \eta \cdot \zeta \text{ for every } \zeta \in C(t), \zeta \neq z(t, \eta) \quad (12)$$

Consider the function

$$\begin{aligned} \hat{f}(t, \eta; x_0) &= \eta \cdot [z(t, \eta) - x^*] \\ &= \int_0^t [\psi(s, \eta) B(s) \hat{u}(s, \eta) + \eta^{n+1} \phi(\hat{u}(s, \eta))] ds - \eta \cdot x^* \end{aligned} \quad (13)$$

where we only consider those  $\eta = (\eta, \eta^{n+1})$  with  $\eta^{n+1} \leq 0$  and  $\eta \neq 0$ . Since the integrand in the above expression is non-negative,  $\hat{f}$  is a continuous, non-decreasing function of  $t$  (for fixed  $\eta$ ). Let us restrict the domain of  $\hat{f}$  to those  $\eta$  for which  $\eta \cdot x^* = -\hat{f}(0, \eta; x_0) > 0$ . If  $z(t^*, \eta) \neq x^*$ , then  $\hat{f}(t^*, \eta; x_0) > 0$  by eqn (12), and consequently,  $\hat{f}(t, \eta; x_0) = 0$  for some  $t \in (0, t^*)$ . On the other hand, suppose that  $z(t^*, \eta) = x^*$ . Then, clearly,  $\hat{f}(t^*, \eta; x_0) = 0$ . Furthermore, it can be shown that in this case  $\hat{f}(t, \eta; x_0) < 0$  for  $t < t^*$ . Indeed, suppose that  $\hat{f}(t', \eta; x_0) = \hat{f}(t^*, \eta; x_0) = 0$ , where  $t' < t^*$ . Then,  $\eta \cdot z(t', \eta) = \eta \cdot z(t^*, \eta)$ , or

$$\int_{t'}^{t^*} [\psi(s, \eta) B(s) \hat{u}(s, \eta) + \eta^{n+1} \phi(\hat{u}(s, \eta))] ds = 0$$

But  $\hat{u}(t, \eta)$  is defined by eqn (11), and this maximum (which is the integrand above) is non-negative, and therefore zero in  $(t', t^*)$ . Since  $\varphi(0) = 0$ ,  $\hat{u}(t, \eta) = 0$  for  $t' < t < t^*$  by virtue of the generalized unique maximum condition. Thus,  $z(t', \eta) = z(t^*, \eta) = x^*$ , which implies that  $\hat{u}(t, \eta)$  is an admissible control which transfers  $x_0$  to 0 in time  $t' < t^*$  with effort  $M$ , contradicting the definition of  $t^*$ .

Now let  $\hat{F}(\eta; x_0)$  be the smallest value of  $t$  for which  $\hat{f}(t, \eta; x_0) = 0$ . Then  $\hat{F}$  is defined, and  $0 < \hat{F}(\eta; x_0) \leq t^*$ . Furthermore,  $\hat{F}(\eta; x_0) = t^*$  if, and only if,  $z(t^*, \eta) = x^*$ , in which case

$\hat{u}(t, \eta)$  is the time-optimal control which transfers  $x_0$  to 0 in time  $t^*$  with effort  $M$ . Again, the vector  $\eta$  which determines the optimal control is not unique, although the optimal control is unique *a.e.*

Thus, summarizing: given a control system described by eqn (1); a convex, closed control region  $U$  which contains the origin; a real valued, continuous function  $\phi(u)$  defined on  $U$ , which is bounded from below and takes on the value 0 at  $u = 0$ ; a control effort constraint of the form  $\int_0^t \phi(u(s)) ds \leq M$ ; and an initial point  $x_0$  in phase space. If (1) the generalized unique maximum condition is satisfied; (2) the sets  $C(t)$  are closed and convex; and (3)  $x \in C(\tau)$  for all  $\tau > t$  implies that  $x \in C(t)$ , then, if there is an admissible control satisfying the effort constraint which transfers  $x_0$  to 0, there is also a unique time-optimal control  $\hat{u}(t, \eta)$  defined by eqn (11), where  $\eta = (\eta, \eta^{n+1})$ , and  $\psi(t, \eta)$  is the solution of eqn (2) with  $\psi(0) = -\eta$ , and  $\eta$  is any vector which maximizes the function  $\hat{F}(\eta; x_0)$  defined as the smallest value of  $t$  for which  $\hat{f}(t, \eta; x_0)$  [see eqn (13)] vanishes.

### III. Minimum Control Effort Optimizations

Now consider the problem in which two points,  $x_0$  and  $x_1$ , in phase space, as well as a time  $T > 0$ , are given, and it is desired to find an admissible control  $u(t)$  which transfers  $x_0$  to  $x_1$  in time  $T$  while minimizing the control effort  $\varepsilon(u(t))$ . Assume that (1)  $U$  is closed and convex; (2)  $\phi(u)$  is continuous and bounded from below on  $U$ ; (3)  $C(T)$  is closed and convex; and (4) the generalized unique maximum condition is satisfied. Finally, assume that the optimization problem is not vacuous, i.e., that there exist admissible controls which transfer  $x_0$  to  $x_1$  in time  $T$ , and that there is more than one possible control effort for such controls.

It follows from (3) that there is an admissible control  $u(t)$  which transfers  $x_0$  to  $x_1$  in time  $T$  with effort  $\lambda$  if, and only if,  $(y, \lambda) \in C(T)$  where  $y = x_0 - X^{-1}(T)x_1$ . Since  $\phi$  is bounded from below and  $C(T)$  is closed, there is a point  $y^* = (y, \lambda^*) \in C(T)$  such that  $(y, \lambda) \in C(T)$  implies that  $\lambda \geq \lambda^*$ ; i.e., there exists an optimal control (with minimum effort  $\lambda^*$ ). Clearly,  $y^*$  is on the boundary of  $C(T)$ , and, by hypothesis,  $C(T)$  contains points  $(y, \lambda)$  with  $\lambda > \lambda^*$ . Since  $y^*$  is an extremal point of  $C(T)$ , there is a support plane to  $C(T)$  at  $y^*$  with normal  $\eta_* = (\eta_*, \eta_*^{n+1})$ , where  $\eta_*^{n+1} < 0$ . It follows as in Section II that  $y^* = z(T, \eta_*)$ , and that  $\hat{u}(t, \eta_*)$  is the unique optimal control.

We shall again construct a function of  $\eta$  which takes on its maximum at those values  $\eta_*$  for which  $\hat{u}(t, \eta_*)$  is the optimal control. Without loss of generality, we shall confine ourselves to those vectors  $\eta$  for which  $\eta^{n+1} = -1$ .

If  $\eta = (\eta, -1)$ , let

$$g(\eta) = \eta \cdot y - \eta \cdot z(T, \eta) \quad (14)$$

If  $z(T, \eta) \neq y^*$ , then

$$\begin{aligned} g(\eta) &= \eta \cdot y - \lambda^* - \eta \cdot z(T, \eta) + \lambda^* \\ &= \eta \cdot y^* - \eta \cdot z(T, \eta) + \lambda^* < \lambda^* \end{aligned}$$

by eqn (12). On the other hand, if  $z(T, \eta) = y^*$ , then  $\lambda^* = g(\eta)$ . Thus,  $g(\eta) \leq \lambda^*$ , and  $g(\eta) = \lambda^*$  if, and only if,  $z(T, \eta) = y^*$  in which case  $\hat{u}(t, \eta)$  is the optimal control which transfer  $x_0$  to  $x_1$  in time  $T$  with minimum effort  $\lambda^*$ .

To summarize: given a control system described by eqn (1), and a time  $T > 0$ ; a convex, closed control region  $U$ ; a real valued, continuous function  $\phi(u)$  which is defined on  $U$  and is bounded from below; control effort defined by  $\int_0^T \phi(u(t)) dt$ ; and an initial point  $x_0$  and a target point  $x_1$  in phase space. If the generalized unique maximum condition is satisfied, and if  $C(T)$  is closed and convex, then (provided the optimization problem is not vacuous) there is a unique (minimum effort) optimal control  $\hat{u}(t, \eta)$  defined by eqn (11), where  $\eta = (\eta - 1)$  and  $\psi(t, \eta)$  is the solution of eqn (2) with  $\psi(0) = -\eta$ , where  $\eta$  is any vector which maximizes the function  $g(\eta)$  given by eqn (14). Furthermore, the maximum of  $g$  is the minimum effort  $\int_0^T \phi(\hat{u}(t, \eta)) dt$ .

#### IV. Computing the Optimal Control

An iterative procedure for finding the maxima of the functions  $F, \hat{F}$ , and  $g$  will now be described.

First consider the problem described in Section I. Let  $\eta_0$  (the initial estimate for the maximizing  $\eta$ ) be any vector such that  $\eta_0 \cdot x_0 > 0$ . Then, if  $\eta_j$  is the estimate from the  $j$ th iteration, choose  $\eta_{j+1}$  to be

$$\eta_{j+1} = \eta_j - K[z(F(\eta_j), x_0), \eta_j] - x_0 \quad (15)$$

where  $K$  is a sufficiently small positive constant. In fact, it will be shown that  $F(\eta_{j+1}) > F(\eta_j)$  (provided that  $K$  is sufficiently small and  $F(\eta_j) < t^*$ ), and that  $\eta_{j+1} - \eta_j$  has the direction of grad  $F(\eta_j)$  if  $F$  is differentiable (in which case our iteration method becomes that of steepest ascent). The argument  $x_0$  in the functions  $f$  and  $F$  will henceforth be dropped for ease of notation.

Consider  $f(t, \eta)$  as a function of  $\eta$  with  $t$  fixed. It is shown in Section V that

$$\frac{\partial}{\partial \eta^i} [\eta \cdot z(t, \eta)] = z^i(t, \eta), \text{ the } i\text{th coordinate of } z(t, \eta)$$

for every  $t > 0$ . Since  $\partial[\eta \cdot (-x_0)]/\partial \eta^i = -x_0^i$ , it follows from eqn (9) that

$$\text{grad } f(t, \eta) = z(t, \eta) - x_0 \quad (16)$$

If  $F(\eta_j) < t^*$ ,  $z(F(\eta_j), \eta_j) \neq x_0$  (by definition of  $t^*$ ), and  $\text{grad } f(F(\eta_j), \eta_j) \neq 0$ . By definition,  $f(F(\eta_j), \eta_j) = 0$ , and, therefore, by well-known properties of the gradient,

$$f(F(\eta_j), \eta_j - K \text{grad } f(F(\eta_j), \eta_j)) < 0$$

if  $K > 0$  is small enough. Hence, by definition of  $F$ ,

$$F(\eta_j - K \text{grad } f(F(\eta_j), \eta_j)) > F(\eta_j)$$

or (see eqns (16) and (15)),

$$F(\eta_{j+1}) > F(\eta_j)$$

It is shown in Section V that  $z(t, \eta)$  is continuous in  $\eta$ . Also,

$$\begin{aligned} \frac{\partial f}{\partial t} &= \frac{\partial}{\partial t} \left\{ \int_0^t \psi(s, \eta) B(s) u(s, \eta) ds \right\} \\ &= \psi(t, \eta) B(t) u(t, \eta) \\ &= \max_{v \in U} \psi(t, \eta) B(t) v \end{aligned}$$

is easily seen to be continuous in  $t$  and  $\eta$ . Hence, if  $\partial f / \partial t \neq 0$  [and since 0 is an interior point of  $U$ ,  $\partial f / \partial t = 0$  only if  $\psi(t, \eta)$

$B(t) = 0$ ] for  $t = F(\eta_j)$  and  $\eta = \eta_j$ , then, by virtue of the implicit function theorem,  $F(\eta)$  has continuous derivatives at  $\eta = \eta_j$  which are given by

$$\frac{\partial F(\eta_j)}{\partial \eta^i} = - \frac{\partial f(F(\eta_j), \eta_j)}{\partial \eta^i} / \frac{\partial f(F(\eta_j), \eta_j)}{\partial t}$$

or,

$$\begin{aligned} \text{grad } F(\eta_j) &= - \left[ \frac{\partial f(F(\eta_j), \eta_j)}{\partial t} \right]^{-1} \text{grad } f(F(\eta_j), \eta_j) \\ &= - \left[ \frac{\partial f}{\partial t} \right]^{-1} [z(F(\eta_j), \eta_j) - x_0] \end{aligned} \quad (17)$$

Thus, the iteration described by eqn (15) is indeed in the direction of 'steepest ascent'.

It is clear from eqn (17) that  $F(\eta)$  has no local extrema at values of  $\eta$  other than those for which  $z(F(\eta), \eta) = x_0$ , and at these values  $F(\eta) = t^*$  and  $u(t, \eta)$  is the time-optimal control. Hence, if the iteration method converges, it must converge to a value of  $\eta$  such that  $u(t, \eta)$  is the time-optimal control.

The computation procedure is particularly adaptable to hybrid analogue digital computers. Typically, it might proceed as follows. A guess for  $\eta$  is made, this value is stored, and  $-\eta$  is used as an initial condition in eqn (2). An analogue computer is then used to compute  $\psi(t, \eta)$  as a function of time. Simultaneously,  $u(t, \eta)$  can be computed with an analogue 'maximizing circuit' which must continuously compute the maximum in  $U$  of a time varying linear function of  $u$  [as given by eqn (6)]. A description of such a circuit, if  $U$  is a polyhedron, is given by Pontryagin et al.<sup>1</sup> A more general computation procedure is described by Neustadt<sup>6</sup>. Simultaneously with  $u(t, \eta)$ ,  $z(t, \eta)$  is obtained from an analogue integration [see eqn (7)], where  $X^{-1}(t)$  is computed as the solution of the equation

$$\frac{d}{dt}(X^{-1}) = -X^{-1}A, X^{-1}(0) = I$$

Finally, the dot product  $\eta \cdot [z(t, \eta) - x_0]$  is computed as an increasing function of time, and the computation is halted when the value of this product is zero. The corresponding time is  $F(\eta)$ . The vector  $[z(F(\eta), \eta) - x_0]$  is then read out and multiplied by an appropriate negative constant  $-K$ , and this value is added to the stored value of  $\eta$  to yield the next approximation, at which point the next iteration is begun.

The maximum of the function  $\hat{F}(\eta)$  of Section II can be found in an almost identical manner. The problem is somewhat more difficult than that described above because of the following factors: (1) the vector  $\eta$  has  $(n+1)$  coordinates (rather than  $n$ ); (2) whereas  $F(\eta)$  is continuous,  $\hat{F}(\eta)$  may have discontinuities, and (3) the computation of  $\hat{u}(t, \eta)$  from eqn (11) is complicated by the fact that the function (of  $v$ ) to be maximized is non-linear.

Finally, the maximum of  $g(\eta)$  (which must be computed to find the optimal control described in Section III) can be found more directly than the maximum of  $F(\eta)$ . We show in Section V that

$$\begin{aligned} \frac{\partial}{\partial \eta^i} \eta \cdot z(T, \eta) &= z^i(T, \eta), \text{ the } i\text{th coordinate of } z(T, \eta), \\ & \quad i = 1, \dots, n \end{aligned}$$

Further,  $\partial(\eta y)/\partial \eta^i = y^i$ , so that (see eqn (14)),

$$\text{grad } g(\eta) = y - \hat{z}(T, \eta)$$

where  $z(T, \eta) = (\hat{z}(T, \eta), z^{n+1}(T, \eta))$ ; i.e.,

$$\hat{z}(T, \eta) = - \int_0^T X^{-1}(t) B(t) \hat{u}(t, \eta) dt$$

Also, note that, by eqn (3) and the definition of  $y$ ,

$$y - \hat{z}(T, \eta) = X^{-1}(T) [x(T, \eta) - x_1]$$

where  $x(t, \eta)$  is the solution of eqn (1) with  $u(t) = \hat{u}(t, \eta)$ .

Thus, an iteration procedure based on the method of steepest ascent with

$$\begin{aligned} \eta_{j+1} &= \eta_j + K \text{grad } g(\eta_j) \\ &= \eta_j + K X^{-1}(T) [x(T, \eta_j) - x_1] \end{aligned}$$

can be used to find the maximum of  $g$ . Here  $K$  is a sufficiently small positive constant, and  $\eta_j = (\eta_j, -1)$ .

## V. Final Derivations and Generalizations

It will be proved that  $z(t, \eta)$  and  $z(t, \eta)$  are continuous functions of  $\eta$  (or  $\eta$ ), and that

$$\frac{\partial}{\partial \eta^i} (\eta \cdot z(t, \eta)) = z^i(t, \eta), \quad \frac{\partial}{\partial \eta^i} (\eta \cdot z(t, \eta)) = z^i(t, \eta)$$

These conclusions follow from general properties of convex sets. Indeed, suppose that  $C$  is any closed and convex set in an  $m$  dimensional vector space and that  $\eta_0$  is a vector in this space with the following property: there exists a neighbourhood  $N$  of  $\eta_0$  such that for every  $\eta \in N$ , the support plane to  $C$ , whose normal (which is directed away from  $C$ ) is  $\eta$ , meets  $C$  at only one point. This point is denoted by  $z(\eta)$ . Then  $z(\eta)$  is continuous at  $\eta = \eta_0$ , and the scalar function  $\phi(\eta) = \eta \cdot z(\eta)$  has continuous derivatives at  $\eta = \eta_0$  given by

$$\frac{\partial \phi(\eta_0)}{\partial \eta^i} = z^i(\eta_0), \quad \text{the } i\text{th coordinates of } z(\eta_0); \quad i=1, \dots, m \quad (18)$$

To prove that  $z(\eta)$  is continuous at  $\eta = \eta_0$ , let  $\eta_j$  be a sequence of vectors such that  $\lim_{j \rightarrow \infty} \eta_j = \eta_0$ ,  $\lim_{j \rightarrow \infty} z(\eta_j) = \zeta$ . Then, by definition of  $z(\eta)$ ,

$$\eta_j \cdot z(\eta_j) \geq \eta_j \cdot z(\eta_0)$$

Passing to the limit as  $j \rightarrow \infty$ , one obtains

$$\eta_0 \cdot \zeta \geq \eta_0 \cdot z(\eta_0)$$

But by hypothesis, this is only possible if  $\zeta = z(\eta)$ . This implies that  $z(\eta)$  is continuous, provided it can be shown that the set of vectors  $z(\eta_j)$  is bounded for any sequence  $\eta_j$  which converges to  $\eta_0$ . But the latter statement is easily verified.

Relation (18) will now be proved. If  $\eta \in N$ ,

$$\begin{aligned} \phi(\eta) - \phi(\eta_0) &= \eta \cdot z(\eta) - \eta_0 \cdot z(\eta_0) = \eta_0 \cdot z(\eta) - \eta_0 \cdot z(\eta_0) \\ &\quad + (\eta - \eta_0) \cdot z(\eta_0) + (\eta - \eta_0) \cdot [z(\eta) - z(\eta_0)] \end{aligned} \quad (19)$$

Let  $\eta - \eta_0 = (0, \dots, 0, \Delta_i \eta, 0, \dots, 0)$ , where  $\Delta_i \eta$  is the  $i$ th coordinate of  $\eta - \eta_0$ . Thus,  $\|\eta - \eta_0\| = |\Delta_i \eta|$ , and  $(\eta - \eta_0) \cdot z(\eta_0) = (\Delta_i \eta) z^i(\eta_0)$ . Hence, it follows from eqn (19) that

$$\left| \frac{\phi(\eta) - \phi(\eta_0)}{\Delta_i \eta} - z^i(\eta_0) \right| \leq \left| \frac{\eta_0 \cdot z(\eta) - \eta_0 \cdot z(\eta_0)}{\Delta_i \eta} \right| + \|z(\eta) - z(\eta_0)\| \quad (20)$$

But  $\eta_0 \cdot z(\eta_0) \geq \eta_0 \cdot z(\eta)$ , and  $\eta \cdot z(\eta) \geq \eta \cdot z(\eta_0)$ , so that

$$0 \geq \eta_0 \cdot z(\eta) - \eta_0 \cdot z(\eta_0) \geq [\eta - \eta_0] \cdot [z(\eta_0) - z(\eta)]$$

or,

$$\frac{|\eta_0 \cdot z(\eta) - \eta_0 \cdot z(\eta_0)|}{|\Delta_i \eta|} \leq \frac{\|\eta - \eta_0\| \|z(\eta_0) - z(\eta)\|}{|\Delta_i \eta|} = \|z(\eta) - z(\eta_0)\|$$

so that, by eqn (20),

$$\left| \frac{\phi(\eta) - \phi(\eta_0)}{\Delta_i \eta} - z^i(\eta_0) \right| \leq 2 \|z(\eta) - z(\eta_0)\|$$

Thus, eqn (18) follows from the continuity of  $z(\eta)$  at  $\eta_0$ .

Finally, mention is made of three problems which are related to the ones described in this paper, which can be treated by techniques similar to those described here.

The first problem is a generalization in which the point  $x_1$  is replaced by an arbitrary compact, convex set. The second problem is the same as that of Section III, except that  $\varepsilon(u(t))$  is redefined by

$$\varepsilon(u(t)) = \sup_{0 \leq t \leq T} \max_j |u^j(t)|$$

This problem was first studied by Krasovskii<sup>7</sup> in his investigation of the time-optimal problem. The third problem is one considered by Ho<sup>8</sup> in his investigation of the same problem. Namely, for given  $x_0$ ,  $x_1$ , and  $T > 0$ , find an admissible control  $u(t)$  which brings  $x_0$  'as close as possible' to  $x_1$  in time  $T$ ; i.e., find a control for which  $\|x(T) - x_1\|$  is minimized, where  $x(t)$  is the solution of eqn (1).

A detailed treatment of these problems will be given elsewhere.

## VI. Computational Studies

A series of computational studies has been carried out to determine the general character of the iterative procedure. These studies were carried out for the time-optimal problem described in Section I. A hybrid analogue digital computer was employed to study second-order systems. An IBM 7090 digital computer was used in the investigation of a third-order system.

A geometrical picture of the iteration procedure for a second-order system is helpful in interpreting the behaviour in the higher-order cases. Figure 1 shows a phase trajectory of  $z(t, \eta)$

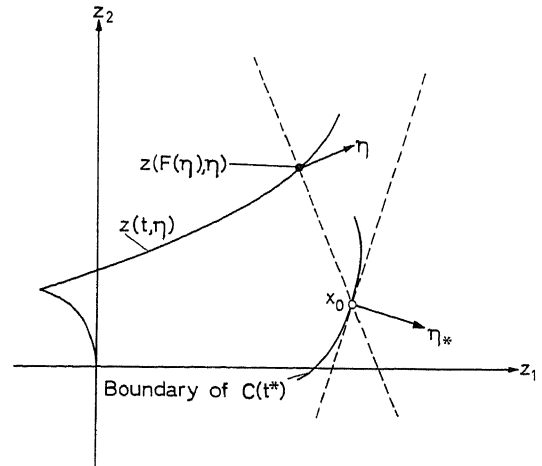


Figure 1. Typical phase trajectory of  $z(t, \eta)$



$z(t, \eta)$  as a function of  $t$  for fixed  $\eta$ . The point  $x_0$  is indicated together with a portion of the 'plane' (line) through  $x_0$ , normal to  $\eta$ . Figure 2 illustrates the procedure for obtaining the gradient for the steepest ascent corrections to  $\eta$  from  $z(t, \eta)$  and  $x_0$ .

The initial tests were conducted for time-optimal regulation ( $x_1 = 0$ ) of a second-order, undamped oscillatory system, with one control:

$$\begin{aligned}\dot{x}_1 &= x_2 \\ \dot{x}_2 &= -\omega^2 x_1 + u(t), |u| \leq 1\end{aligned}\quad (21)$$

The initial direction of  $\eta$  was taken opposite the state vector, i.e.,

$$\eta_1 = -\frac{x_0}{\|x_0\|}$$

The equations for  $z(t, \eta_i)$  were integrated on the analogue portion of the hybrid computer. The digital computer stopped the integration at the time  $t = F(\eta_i)$  when  $f(t, \eta_i; x_0) = 0$ . The corrections to  $\eta_i$  were obtained in the digital computer, fed to the analogue and the process repeated. The cycle continued until the terminal point of the  $z$  trajectory was within a specified distance of  $x_0$ .

The numbered sequence of  $z$  trajectories shown in Figure 3

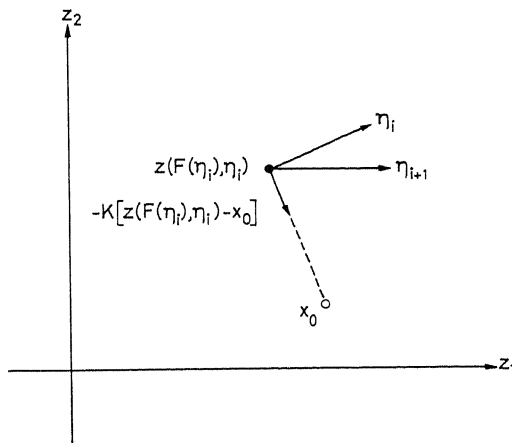


Figure 2. Illustration of gradient procedure

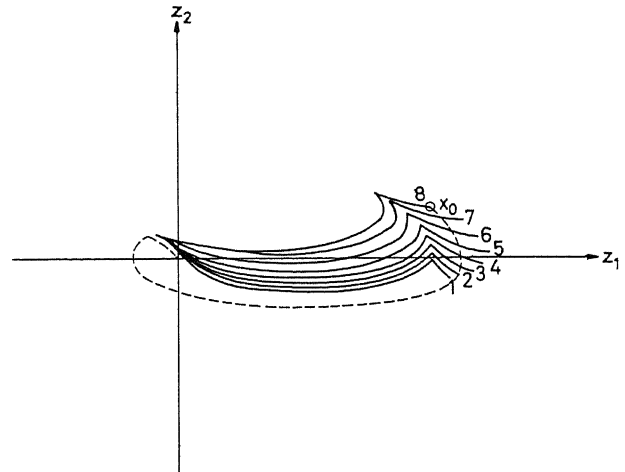


Figure 3. Convergence of  $z$  trajectories, Case 1

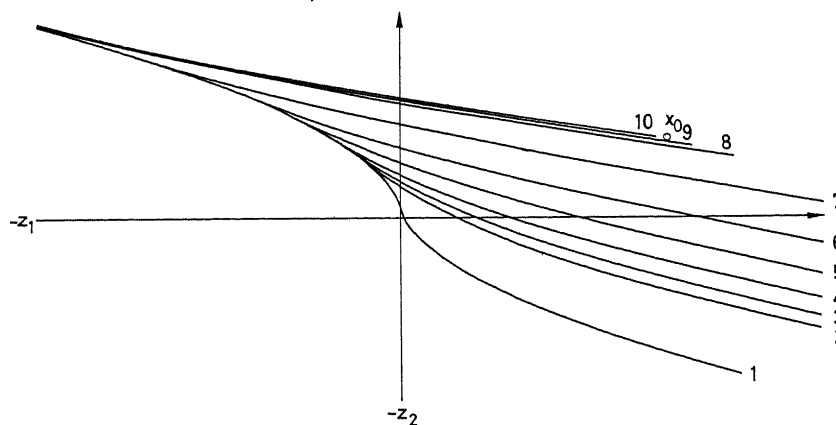


Figure 4. Convergence of  $z$  trajectories, Case 2

converges to  $x_0$  in eight trials. The broken line represents the phase portrait of the system responding to the optimal control.

A proper choice for the constant  $K$  in eqn (15) is very important. In the problem under discussion, the value  $K = 1$  proved to be satisfactory. However, it was possible to cause large oscillations about the correct solution by choosing  $K \gg 1$ .

The second case studied was time-optimal regulation of the system  $\dot{x}_1 = x_2, \dot{x}_2 = u(t), |u| \leq 1$ .

A set of  $z(t, \eta)$  trajectories for this system is shown in Figure 4. The support plane to  $C(t^*)$  at  $x_0$  is almost parallel to the  $z(t, \eta_*)$  trajectory. In this case, a small change in  $\eta$  produces a large displacement in  $z(F(\eta), \eta)$ . Hence, it was necessary to programme  $K$  as a function of  $\|z(t, \eta) - x_0\|$  in order to avoid large oscillations, and, on the other hand, to keep the number of iterations small. In the case shown, the value  $K = 0.1$  proved satisfactory near the solution, but larger values were employed for large values of  $\|z(t, \eta) - x_0\|$ .

In addition to studies of time-optimal regulation, a number of tests on second-order systems were made to determine the behaviour when 'pursuing' moving target points. Satisfactory results were obtained in these tests also.

The application of this technique to second-order systems presented no particular difficulty from the standpoint of stability of the iterations or finding a proper choice for  $K$ . The study of third-order systems, however, led to some interesting difficulties.

A simplified steering problem provided this example. The system equations are:

$$\dot{x}_1 = \frac{c}{\mu - t} u_1$$

$$\dot{x}_2 = x_3$$

$$\dot{x}_3 = \frac{c}{\mu - t} u_2 - 30$$

( $c$  and  $\mu$  are given constants), and  $U$  is given by  $(u_1)^2 + (u_2)^2 \leq 1$ . Here, also,  $x_1 \neq 0$ .

This problem was solved entirely on a digital computer. Again,  $z(F(\eta), \eta)$  proved to be an extremely sensitive function of  $\eta$ . As a result, it was necessary to determine the time  $F(\eta)$  with great precision. Indeed, the maximum error which could

be tolerated was of the order of  $10^{-4}$  sec. This sensitivity is illustrated in Table 1, whose values were chosen from a representative case. This table presents  $F(\eta)$  and the three components of  $[z(F(\eta), \eta) - x_0]$  for four representative values of  $\eta$ .

The extreme sensitivity of  $z(F(\eta), \eta)$  made it very difficult to converge to the correct value of  $\eta$ , or correspondingly to bring  $z(F(\eta), \eta)$  near  $x_0$ . In order to circumvent this difficulty the following two steps were taken. First, a carefully chosen programme for  $K$  was adopted. Secondly when successive values of  $F(\eta_i)$  differed by a preselected amount (corresponding to  $F(\eta_i)$  being very near the minimum time), a different computational scheme was used to obtain a precise estimate for  $\eta_*$ . When the refinements were carried out, the convergence was rapid, with the number of trials generally less than 50. These computational aspects are discussed in more detail elsewhere.

Table 1. Typical Variation of  $F$  and  $z$  with  $\eta$

Components of $\eta$			$F(\eta)$	Components of $z(F(\eta), \eta) - x_0$		
0.54580027	0.01316405	1.0729829	148.56863	0.830	22,700	-7.373
0.54580027	0.01316405	1.0729839	148.56906	0.397	52,705	-16.369
0.54580027	0.01316405	1.0734830	148.56824	1.259	7,410	1.640
0.54580027	0.01314405	1.0734830	148.56763	10.711	-121,610	31.768

## References

- PONTRYAGIN, L. S., BOLTYANSKII, V. G., GAMKRELDIDZE, R. V. and MISHCHENKO, E. F. The mathematical theory of optimal processes. (English transl.) 1962. New York; Interscience
- NEUSTADT, L. W. Synthesizing time optimal control systems. *J. Math. Anal. Applic.* 1, No. 4 (1960), 484
- BELLMAN, R., GLICKSBERG, I. and GROSS, O. On the 'bang-bang' control problem. *Quart. Appl. Math.* 14, No. 1 (1956), 11
- LASALLE, J. P. The time optimal control problem. *Contributions to the Theory of Nonlinear Oscillations*, Vol. 5, pp. 1-24. 1960. Princeton; Princeton University Press
- NEUSTADT, L. W. Time optimal systems with position and integral limits. *J. Math. Anal. Applic.* 3, No. 3 (1961), 406
- NEUSTADT, L. W. Applications of linear and nonlinear programming techniques. *Proc. Third Internat. Conf. Analog Computation*. Opatija, Yugoslavia (1961) 197
- KRASOVSKII, N. N. On the theory of optimal control. *Avtomat. Telemekh., Moscow*, 18, No. 11 (1957) 960
- HO, Y. C. A successive approximation technique for optimal control systems subject to input saturation. *J. Bas. Engng* 84, Ser. D, No. 1 (1962), 33

## DISCUSSION

R. SARGENT, *Imperial College, London*

I was interested in Dr. Neustadt's comment that the proof of convexity of the set  $C(t)$  in the paper could be extended to some non-linear systems. The usual proof depends heavily on the linearity of the equations; but there is an additional difficulty if the equations are non-linear, since the set is defined in terms of the matrizant  $X(t)$ , which, of course, is only defined for a linear system. One can still write the solution in the same form as eqns (3) for a non-linear system, but then  $X$  is also a function of the initial conditions, so that we no longer obtain a set with the properties of  $C(t)$ , and it is not clear that the convexity proof provides the same type of solution as for the linear case.

L. W. NEUSTADT, *in reply*

Professor Sargent misunderstood my remarks concerning the convexity of the sets  $C(t)$  under conditions less stringent than those described in our paper.

The sets  $C(t)$  are convex if the equations are non-linear in  $u$  but linear in  $x$ ; i.e. if the control system is described by an equation of the form

$$\dot{x}(t) = A(t)x(t) + \phi(u/t)$$

where  $\phi$  is a continuous function, and if the set  $U$  of admissible values for the control variable  $u$  is compact (but not necessarily convex).

The proof of the convexity and compactness of the sets  $C(t)$  under the above hypotheses is given in another paper<sup>1</sup>.

## Reference

- NEUSTADT, L. W. The existence of optimal controls in the absence of convexity conditions. *J. Math. Anal. Applications* 7 (1963)
- A. R. M. NOTON, *Electrical Engineering Department, University of Nottingham, England*

In synthesizing optimal controls, the author has pointed out that the calculations reduce to the solution of differential equations with mixed boundary conditions. Because of the difficulties in coping with such boundary conditions, he reformulated the problem so that it becomes one of maximizing a function of the initial conditions  $\psi(0)$ , eqn (2). Now, although the second-order system quoted, eqn (21), presented no difficulties, with the third-order system, the  $z$  function was extremely sensitive to variations in  $\eta$  [ $= -\psi(0)$ ]. An entirely digital computation then became necessary.

This difficulty should be compared with the experiences of Westcott, Florentin and Pearson, who adopted a more direct approach, and in their paper have suggested a procedure for coping with the sensitivity to variations in  $\psi(0)$ . The author has not. The benefits of his mathematical reformulation are not therefore evident. Perhaps he will discuss them.

L. W. NEUSTADT, *in reply*

The sensitivity of  $z(t, \eta)$  to small changes in  $\eta$  is indeed a serious, but not insurmountable, problem in applying the computational method described in our paper.

The method of circumventing this difficulty, as discussed by Westcott *et al.*, is confined to the *free* right-hand endpoint problem, whilst in our paper, the *fixed* right-hand endpoint problem is the one under consideration. In addition, the derivation given by Westcott *et al.* is a formal one, which can run into difficulties when the derivatives  $\partial\psi/\partial p_i$  and  $\partial\eta/\partial p_i$  fail to exist (and such cases are not difficult to run into).

J. A. STILES, *Trinity College, Cambridge University, Cambridge, England*

Does the author consider that the results of Section I could be extended to simple non-normal systems?

L. W. NEUSTADT, *in reply*

The proofs in Section I lean rather heavily on the so-called unique maximum condition, or, in other words, on the normality of the control system. Whether the proofs can be extended to non-normal systems is open to conjecture.

It is clear that certain difficulties will arise. Geometrically, non-normality generally implies that the sets  $C(t)$  have flat portions, i.e. that a support plane to  $C(t)$  meets  $C(t)$  at more than one point. For this reason, the geometric arguments given in Section V must be modified. Analytically, non-normality means that the optimal control is not completely determined by the proper  $\psi(0)$ .

If the unique maximum condition is satisfied by the optimal control being sought, one should expect that the difficulties described in the preceding paragraph can be circumvented even in non-normal systems.

R. KULIKOWSKI, *Polish Academy of Sciences, Warsaw U 1. Koszykowa 75/18, Poland*

The problem considered in this paper belongs to the wide and important class of multi-constrained control problems. A number of different analytical methods can be applied for the determination of the optimum control signal  $u(t)$ , which is constrained and minimizes the given functional  $F[u]$ . A pretty general method may be based on the calculus of variations in an abstract space. In the case of strongly differentiable functionals  $F[u]$ , the necessary condition for optimality follows from the known Lusternik theorem for conditional variational problems. The sufficient condition of optimality for continuous (or at

least weakly semi-continuous) functionals  $F[u]$ , which are given in a region  $w$  of a weakly compact Banach space, can be determined by the so-called Weierstrass theorem<sup>1</sup>. When  $\|\text{grad } F[u]\| > 0$ , the optimum control  $n$  belongs to the boundary of  $w$  and the synthesis of the optimum control system can be reduced to the solution of a number of equations for the unknown switching instants. However, this method requires familiarity with some abstract concepts of non-linear functional analysis, which are not very popular among engineers. A great advantage of the geometric arguments used by Dr. Neustadt and others in their papers is in the simplicity of the physical interpretation. However, the computational scheme used for determination of the unknown coefficients does not seem to be so simple. For example, in order to find the best time optimal control with amplitude constraints  $|u(t)| \leq 1$  for the second-order system:  $\dot{x}_1 = x_2$ ,  $\dot{x}_2 = u(t)$ , a number of trajectories should be constructed (as shown in Figure 4). On the other hand, using the known method, based on the construction of the switching curve in phase-plane, the optimum controller can be synthesized very easily.

In connection with this problem it is interesting to know whether a faster convergence of the iterations (for  $\eta_i$  coefficients) could be obtained if the  $K$  coefficient were determined from the condition

$$\frac{dF}{dK} \{ \eta_i - K \text{grad } f[F(\eta_i), \eta_i] \} = 0, \quad i = 1, 2$$

or chosen in another 'optimal' way.

#### Reference

- <sup>1</sup> WEINBERG, M. M. *Variational Methods of Investigation of Non-linear Operators* (in Russian). 1956. Moscow; Gittl

L. W. NEUSTADT, *in reply*

Being unfamiliar with the calculus of variations in abstract spaces, I cannot comment in detail on the early part of Professor Kulikowski's discussion. However, I should like to point out that the basic aim of our paper was to reduce the problem of finding an optimum function to that of maximizing a function of a finite number (equal to the order of the basic differential equation) of variables, since the latter seems so attractive computationally. Admittedly, even the second problem is not trivial. The methods of steepest ascent suggested in our paper is the most naive, and does result in computational difficulties.

Since the writing of our paper, computer studies using other, more refined hill climbing techniques (in particular, a method described recently in the British Computer Journal by Powell) have been carried out, and time-optimal trajectories have been computed with relative ease. These results will be reported by Dr. Paiewonsky.

As pointed out by Professor Kulikowski, the second-order examples discussed in Section VI are easily synthesized by known switching curves. However, in higher-order systems where switching surfaces are difficult to obtain and to simulate, or in systems where the optimal control can take on a large number of values, the method described in the paper should offer some strong advantages.

With respect to the last of Professor Kulikowski's points, Powell's hill climbing method specifies the value of  $K$  in eqn (15).

# An Application of Optimal Control to Midcourse Guidance

J. S. MEDITCH and L. W. NEUSTADT

## Summary

A general midcourse guidance problem is formulated as an optimal control problem. The problem consists of bringing a vehicle on to a nominal trajectory with a minimum expenditure of fuel. An amplitude constraint is placed on the thrust (or control) vector. The equations are linearized by considering perturbations about the nominal.

The form of the optimal control follows from the maximum principle and is a vector function of the system adjoint solution. The optimal control is 'on' only when a scalar function of the system adjoint solution exceeds a fixed threshold. When the control is 'on', the length of the control vector is equal to the constraint limit while its 'direction' is governed by the 'direction' of the system adjoint solution.

The usually difficult problem of obtaining the initial conditions for the adjoint solution is reduced to the problem of maximizing a particular function. It is shown that this function possesses a unique maximum and that its gradient assumes an especially simple form. A direct synthesis procedure then follows by applying the method of steepest ascent. The method developed is particularly suited for digital computer solution.

## Sommaire

Ce rapport présente un problème de pilotage de missile en navigation, comme un problème d'optimisation. Il s'agit de mettre le missile sur sa trajectoire nominale en utilisant un minimum de combustible. Une contrainte est appliquée au vecteur de la poussée. Prenant en considération des perturbations par rapport à un régime nominal.

La forme de l'optimisation résulte du principe du maximum et constitue une fonction vectorielle de la solution adjointe du système. Cette optimisation n'est réalisée que lorsque une fonction scalaire de cette solution, dépasse certain seuil. Quand cette condition est remplie, la longueur du vecteur caractéristique du système est égale à celle qui résulte de la contrainte, tandis que sa direction est fixée par celle donnée par la solution adjointe du système. La difficulté habituelle de déterminer les conditions initiales de la solution adjointe du système se réduit au problème de la maximisation d'une fonction particulière. Il est montré que cette fonction ne possède qu'un seul maximum et que son gradient est d'une forme particulièrement simple. Une procédure de synthèse directe en découle en appliquant la méthode de la pente la plus raide. Cette méthode se prête particulièrement bien à un traitement au moyen d'un calculateur numérique.

## Zusammenfassung

Ein allgemeines Problem der Lenkung während der mittleren Flugphase (Startphase—mittlere Phase—Endphase) wird als optimales Regelproblem formuliert. Die Schwierigkeit besteht darin, den Flugkörper bei geringstem Brennstoffverbrauch auf eine vorgeschriebene Flugbahn zu bringen. Der Schubvektor (oder der Vektor der Regelung) unterliegt einer Amplitudenbegrenzung. Die Linearisierung der Gleichungen geschieht dadurch, daß man Störungen um die Nennwerte annimmt.

Die Art der Optimalwertregelung folgt aus dem Maximumprinzip und ist eine Vektorfunktion der adjungierten Lösung des Systems. Die optimale Regelung wird nur dann wirksam, wenn eine Skalarfunktion der adjungierten Lösung des Systems einen festen Schwellwert über-

schreitet. Bei wirksamer Regelung ist die Länge des Vektors der Regelung gleich der vorgegebenen Begrenzung, während seine „Richtung“ von der „Richtung“ der adjungierten Lösung des Systems abhängt.

Die normalerweise schwierige Gewinnung der Anfangsbedingungen für die adjungierte Lösung wird auf die Maximierung einer bestimmten Funktion zurückgeführt. Es zeigt sich, daß diese Funktion ein eindeutiges Maximum besitzt und daß ihr Gradient eine besonders einfache Form annimmt. Daraus folgt unmittelbar ein Syntheseverfahren nach der Methode des steilsten Anstieges. Die hier entwickelte Methode ist besonders gut für die Lösung auf Digitalrechnern geeignet.

## Introduction

Recent advances in optimal control theory<sup>1-3</sup> have led naturally to an interest in applying these results to practical problems. The need for application studies is especially strong in the field of missile and space vehicle guidance and control, where such factors as fuel consumption, payload, mission time, and target errors are critical. Some results have been obtained in this field<sup>4-7</sup>.

This paper presents a synthesis of the control function which is optimal in the sense of minimum fuel for the midcourse phase of a broad class of space missions. The central result of the paper is the development of an iterative computational procedure for synthesizing the optimal control. The procedure is simple in form and well suited for digital computation.

## Midcourse Guidance as an Optimal Control Problem

A number of space missions may be subsumed under the configuration shown in *Figure 1*. The problem is that of transferring a space vehicle from one moving point *A* to another moving point *B*. Both *A* and *B* may lie on bodies which are spinning as well as moving along their respective orbits.

Typically, the mission and its associated guidance operations are separated into three phases: launch or boost, midcourse, and terminal. These are shown in the figure. If the launch phase

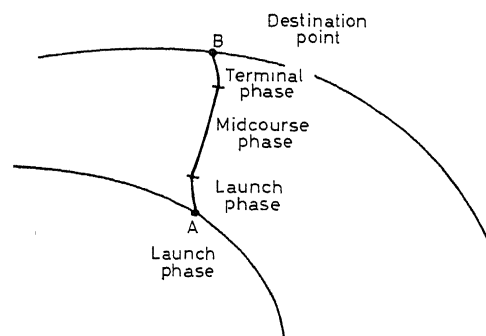


Figure 1. General configuration for space missions

could be executed perfectly, the vehicle would 'free-fall' to the destination point without requiring any corrections. However, imperfect launch guidance, resulting from sensor errors, incorrect thrusting, external disturbances, etc., causes errors to exist at the conclusion of this phase. Hence, midcourse corrections may be required. The term errors is used here to denote position and velocity deviations from the ideal 'free-fall' trajectory which is hereafter termed the nominal trajectory.

There are basically two major problems associated with the overall guidance problem for each phase of a space mission. The first deals with the sensing and processing of information to determine the vehicle's state, such as its position and velocity. The second is concerned with utilizing this information to guide and control the vehicle so that the mission objectives are achieved. This paper presents a solution to the second problem for the midcourse phase, under the assumption that the position and velocity errors are known at the termination of the launch phase. These errors are assumed to be small enough to permit linearization of the vehicle's equations of motion about the nominal trajectory.

In order to allow freedom in the terminal phase for manoeuvring and landing or docking, it is required that the midcourse corrections reduce the errors to zero at the end of the midcourse phase. In addition, it is required that the midcourse phase be executed in a fixed time equal to the time for the nominal flight. Physically, these two requirements mean that the vehicle will arrive at the desired destination point with the same velocity and at the same time as a vehicle following the nominal trajectory. Thus, the task of the terminal phase will be reduced considerably.

It will be assumed that the midcourse phase is to be executed using a minimum amount of fuel in order to maximize the payload. In addition, it is assumed that the corrective thrust magnitude is constrained for obvious practical reasons.

The equations of motion of a space vehicle subject only to gravitational and propulsive forces may be written as

$$\ddot{x}_i = f_i(x_1, x_2, x_3, t) + \frac{1}{m} T_i(t) \quad i=1, 2, 3 \quad (1)$$

where  $x_1, x_2$ , and  $x_3$  are the coordinates of the centre of mass of the vehicle in a Cartesian inertial coordinate system and the dots denote differentiation with respect to time. The gravitational accelerations, as represented by the  $f_i$ , are time dependent since the attracting bodies, e.g. planets, may be moving. The  $T_i(t)$  represent the propulsive forces, and  $m$  is the mass of the vehicle. It is assumed that the mass of the fuel consumed during midcourse flight is small compared to the total vehicle mass. Thus,  $m$  is essentially constant during the mission.

$$\text{Defining} \quad x_{i+3} = \dot{x}_i \quad i=1, 2, 3 \quad (2)$$

the six equations, (1) and (2), may be written in vector form as

$$\dot{x} = f(x, t) + \frac{1}{m} T(t) \quad (3)$$

Here  $x$  is the vector  $(x_1, \dots, x_6)$ ;  $f$  depends only on the first three components of  $x$  (as well as time); and the vector  $T(t)$  is related to the  $T_i(t)$  of (1) in an obvious way.

It is sometimes more advantageous to use polar or spherical coordinates rather than Cartesian coordinates as in (1). In vector form an equation similar to (3) is obtained:

$$\dot{y} = P(y, t) + Q(y) T(t) \quad (4)$$

For the sake of generality, it will be assumed the equations of motion are of the form of (4).

Let  $y = Y + \delta y$  where  $Y$  represents a nominal trajectory satisfying the free-fall equation

$$\dot{Y} = P(Y, t)$$

Since  $y$  satisfies (4),  $\delta y$  satisfies the equation

$$\delta \dot{y} = \left( \frac{\partial P}{\partial y} \right) \delta y + Q(y) T(t) + \dots \quad (5)$$

where  $(\partial P / \partial y)$  is a  $6 \times 6$  matrix of partial derivatives (evaluated along the nominal trajectory  $Y$ ), and the dots on the right denote higher order terms in  $\delta y$ .

The term  $Q(y) T(t)$ , which represents a propulsive correction, will be considered to be small of the first order. Therefore, the difference

$$[Q(Y) - Q(y)] T(t)$$

will be considered to be a 'second-order' term. Thus, neglecting higher order terms, eqn (5) can be written as

$$\delta \dot{y} = \left( \frac{\partial P}{\partial y} \right) \delta y + Q(Y) T(t) \quad (6)$$

Defining

$$x(t) = \delta y(t) \quad A(t) = \left( \frac{\partial P}{\partial y} \right)$$

$$u(t) = T(t) \quad B(t) = Q(Y)$$

(6) becomes

$$\dot{x}(t) = A(t) x(t) + B(t) u(t) \quad (7)$$

As an example consider the two-dimensional, restricted two-body problem. Let  $r_1$  and  $\theta_1$  be the polar coordinates of the centre of mass of the vehicle in an inertial coordinate system with origin in the centre of the attracting body. Let  $u_1$  and  $u_2$  be the components of the vehicle's radial and tangential thrust, respectively (see Figure 2). Letting  $\dot{r}_1 = r_2$  and  $\dot{\theta}_1 = \theta_2$ , it is well known that the equations of motion take the form

$$\frac{d}{dt} \begin{bmatrix} r_1 \\ r_2 \\ \theta_1 \\ \theta_2 \end{bmatrix} = \begin{bmatrix} r_2 \\ r_1 \theta_2^2 - \frac{\mu}{r_1^2} \\ \theta_2 \\ -2 \frac{r_2 \theta_2}{r_1} \end{bmatrix} + \begin{bmatrix} 0 & 0 \\ \frac{1}{m} & 0 \\ 0 & 0 \\ 0 & \frac{1}{mr_1} \end{bmatrix} \begin{bmatrix} u_1 \\ u_2 \end{bmatrix} \quad (8)$$

where  $m$  is the vehicle's mass,  $\mu = GM$  where  $G$  is the universal gravitational constant, and  $M$  is the mass of the attracting body.

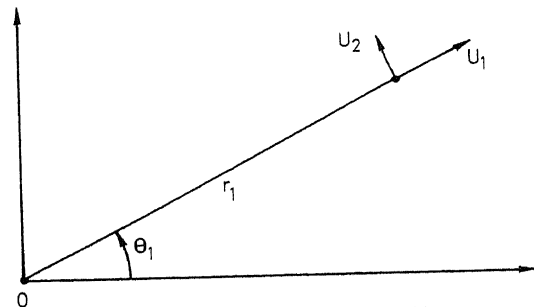


Figure 2. Restricted two-body problem

Linearizing (8), with  $R$ ,  $\Theta$ ,  $\dot{R}$ , and  $\dot{\Theta}$  the values of  $r_1$ ,  $\theta_1$ ,  $r_2$ , and  $\theta_2$  along the nominal trajectory, respectively, yields

$$\frac{d}{dt} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \end{bmatrix} = \begin{bmatrix} 0 & 1 & 0 & 0 \\ \dot{\theta}^2 + \frac{2\mu}{R^3} & 0 & 0 & 2R\dot{\theta} \\ 0 & 0 & 0 & 1 \\ \frac{2\dot{R}\dot{\theta}}{R^2} & -\frac{2\dot{\theta}}{R} & 0 & -\frac{2\dot{R}}{R} \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \end{bmatrix} + \begin{bmatrix} 0 & 0 \\ \frac{1}{m} & 0 \\ 0 & 0 \\ 0 & \frac{1}{mR} \end{bmatrix} \begin{bmatrix} u_1 \\ u_2 \end{bmatrix}$$

where  $x_1 = \delta r_1$ ,  $x_2 = \delta r_2$ ,  $x_3 = \delta \theta_1$ , and  $x_4 = \delta \theta_2$ .

Returning to the general case, it will be assumed hereafter that  $x(t)$  is an  $n$ -dimensional column vector, the state vector;  $u(t)$  is an  $r$ -dimensional column vector, the control vector; and  $A(t)$  and  $B(t)$  are  $n \times n$  and  $n \times r$  matrices, respectively.

Since the time interval over which midcourse guidance is performed is fixed, it is assumed that  $0 \leq t \leq \tau$  in (7), where  $\tau$  is a known constant. The deviations which exist at the initiation of midcourse guidance are denoted by  $x(0)$ . To null these errors, it is required that  $x(\tau) = 0$ .

An amplitude constraint is placed on the thrust or control vector by requiring  $\|u(t)\| \leq \Lambda$ , where  $\Lambda$  represents the maximum thrust available. The symbol  $\| \cdot \|$  denotes the Euclidean norm given by

$$\|u(t)\| = \sqrt{u_1^2(t) + \dots + u_r^2(t)}$$

where the  $u_i(t)$  are the components of  $u(t)$ . Note that  $\|u(t)\|$  is simply the length or magnitude of the control (thrust) vector. Replacing the matrix  $B(t)$  by  $\Lambda B(t)$ , and  $u(t)$  by  $(1/\Lambda)u(t)$ , the constraint can be assumed to be normalized, and of the form  $\|u(t)\| \leq 1$ .

Assuming a constant rate of expulsion of fuel, the amount of fuel consumed during the midcourse phase is proportional to the time integral of the length of the control (thrust) vector. Thus, the cost function to be minimized is given by

$$S = \int_0^\tau \|u(t)\| dt \quad (9)$$

The formal mathematical statement of the problem is now presented. Given the linear dynamical system of (7), it is desired to drive the system from a known initial state  $x(0)$  in phase space to the origin in a fixed time  $\tau$  such that the cost function (9) is minimized subject to the constraint  $\|u(t)\| \leq 1$  for all  $t$ ,  $0 \leq t \leq \tau$ .

Controls for which  $\|u(t)\| \leq 1$  and for which each component of  $u(t)$  is measurable for  $0 \leq t \leq \tau$  are termed *admissible*. The set  $U$  of all such controls is defined by the relation

$$U = [u(t) : \|u(t)\| \leq 1, u(t) \text{ measurable}, 0 \leq t \leq \tau]$$

where  $u(t)$  is an  $r$ -dimensional column vector.

An admissible control which minimizes the integral in (9) and drives the system given by (7) from the initial state  $x(0)$  to the origin in time  $\tau$  is called an *optimal control*.

### Derivation of the Optimal Control

The response of the system given by (7) is

$$x(t) = X(t) \left[ x(0) + \int_0^t X^{-1}(s) B(s) u(s) ds \right] \quad (10)$$

where  $X(t)$  is the  $n \times n$  matrix solution of

$$\dot{X}(t) = A(t) X(t), \quad X(0) = I$$

where  $I$  is the identity matrix.

Since  $X(\tau)$  is non-singular,  $x(\tau) = 0$  if and only if

$$-x(0) = \int_0^\tau X^{-1}(s) B(s) u(s) ds \quad (11)$$

In other words, for an admissible control  $u(t)$ , (11) determines the initial state from which it is possible to reach the origin in time  $\tau$  using the control  $u(t)$ .

Let an additional state variable  $x_{n+1}$  be defined by

$$x_{n+1}(t) = \int_0^t \|u(s)\| ds$$

Clearly,  $x_{n+1}(\tau)$  is the cost function.

Now consider the set  $\Omega(\tau)$  defined by the relation

$$\Omega(\tau) = \left\{ \left[ \int_0^\tau X^{-1}(t) B(t) u(t) dt, \int_0^\tau \|u(t)\| dt \right] : u(t) \text{ admissible} \right\}$$

A point  $(x_1, \dots, x_n, x_{n+1})$  belongs to  $\Omega(\tau)$  if and only if there is an admissible control  $u(t)$  which drives the initial state  $(-x_1, \dots, -x_n)$  to the origin in time  $\tau$  with cost  $x_{n+1}$ .

Since  $-u(t)$  is admissible if  $u(t)$  is, and since  $\|-u(t)\| = \|u(t)\|$ , it is clear from the definition of  $\Omega(\tau)$  that this set is symmetric about the cost ( $x_{n+1}$ ) axis.

It is shown in the Appendix that  $\Omega(\tau)$  is convex, closed, and bounded. A typical three-dimensional representation of  $\Omega(\tau)$  is given in Figure 3 for a second-order system. For higher-order systems the simple two-dimensional representation of Figure 4 will be helpful in the discussions which follow.

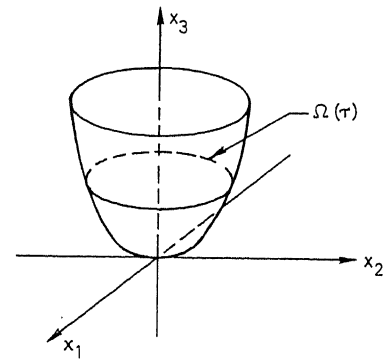


Figure 3. Representation of  $\Omega(\tau)$  for second-order systems

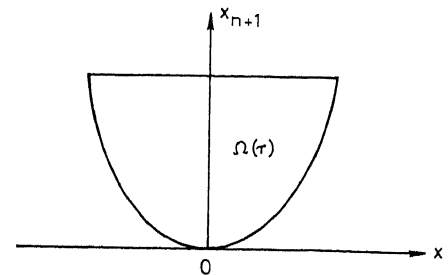


Figure 4. Simplified representation of  $\Omega(\tau)$  for higher-order systems

For some initial states it will be impossible to reach the origin even if full control  $\|u(t)\| = 1$  is utilized throughout the interval  $[0, \tau]$ . The set of all such initial states is termed *degenerate*. Only cases where the initial states are non-degenerate will be considered here.

If a non-degenerate  $x(0)$  is given, the minimum cost to reach the origin in time  $\tau$  is the least  $x_{n+1}(\tau)$  for which  $(-x(0), x_{n+1}(\tau)) \in \Omega(\tau)$ . This cost is denoted by  $x_{n+1}^0(\tau)$ . Obviously,  $(-x(0), x_{n+1}^0(\tau))$  is a boundary point of  $\Omega(\tau)$ .

Let  $y_0 = (-x(0), x_{n+1}^0(\tau))$ . Since  $\Omega(\tau)$  is convex, there exists at least one  $(n+1)$ -dimensional row vector  $\eta^*$  such that

$$\eta^* \cdot y_0 \geq \eta^* \cdot \omega \quad (12)$$

for all  $\omega \in \Omega(\tau)$ . That is, a hyperplane of support may be constructed at  $y_0$  with  $\eta^*$  a vector normal to this hyperplane at  $y_0$ . This result has a simple geometric interpretation which is given in Figure 5. The hyperplane of support is 'tangent' to the boundary of  $\Omega(\tau)$  at  $y_0$ . The vector  $\eta^*$  is normal to this plane and is directed 'out of' or 'away from'  $\Omega(\tau)$ . Observe that the components of this vector are proportional to the direction numbers of the line 'normal' to the boundary of  $\Omega(\tau)$  at  $y_0$ .

It is geometrically obvious (see Figure 5) and easily shown that the  $(n+1)$ st component of  $\eta^*$  is non-positive. With little

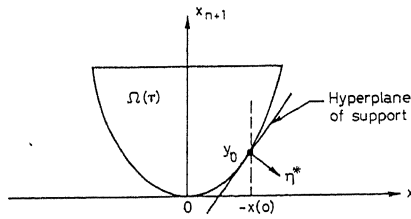


Figure 5. Geometric interpretation of hyperplane of support and normal vector  $\eta^*$

loss of generality, it may be assumed negative. Since the length of  $\eta^*$  is obviously immaterial, the  $(n+1)$ st component of  $\eta^*$  is taken to be  $-1$ . Henceforth, it will be assumed that all 'normal' vectors to  $\Omega(\tau)$  with negative  $(n+1)$ st component have this component equal to  $-1$ .

The function (of  $\omega$ )  $\eta^* \cdot \omega$  attains its maximum<sup>8</sup> in  $\Omega(\tau)$  when  $\omega = y_0$ . In general, for  $\omega \in \Omega(\tau)$  given by

$$\omega = \left[ \int_0^\tau X^{-1}(t) B(t) u(t) dt, \int_0^\tau \|u(t)\| dt \right]$$

$\eta^* \cdot \omega$  becomes

$$\eta^* \cdot \omega = \int_0^\tau [\eta \cdot X^{-1}(t) B(t) u(t) - \|u(t)\|] dt \quad (13)$$

where  $\eta^* = (\eta, -1)$ .

It is desired to select an admissible control  $u(t)$  which maximizes  $\eta^* \cdot \omega$ . This function will be maximized by maximizing the integrand of (13) for all  $t$ ,  $0 \leq t \leq \tau$ .

Let

$$f(t) = \eta \cdot X^{-1}(t) B(t) u(t) - \|u(t)\| \quad (14)$$

and define

$$a(t) = \eta \cdot X^{-1}(t) B(t) \quad (15)$$

which is an  $r$ -dimensional row vector.

Substituting (15) into (14) gives

$$f(t) = a(t) \cdot u(t) - \|u(t)\| \quad (16)$$

as the function to be maximized by the choice of an admissible control  $u(t)$  over the interval  $[0, \tau]$ .

If  $a(t) = 0$ ,  $f(t)$  will clearly be maximized by setting  $u(t) = 0$ . If  $a(t) \neq 0$ ,  $f(t)$  will be maximized only if  $u(t)$  has the same direction as  $a(t)$  in which case

$$a(t) \cdot u(t) = \|a(t)\| \cdot \|u(t)\|$$

and

$$f(t) = \|u(t)\| [\|a(t)\| - 1] \quad (17)$$

The maximization of  $f(t)$  is now separated into the following three cases:

$$(a) \quad \|a(t)\| > 1$$

$$(b) \quad \|a(t)\| < 1$$

$$(c) \quad \|a(t)\| = 1$$

Case (a)

Whenever  $\|a(t)\| > 1$ ,  $f(t)$  is maximized by setting  $\|u(t)\| = 1$ , the maximum permissible. Hence, since the two vectors are in the same direction, the choice of  $u(t)$  is

$$u(t) = \frac{a'(t)}{\|a(t)\|}$$

where the prime denotes the transpose. Note that the components of  $u(t)$  are simply the components of  $a'(t)$  normalized so that  $\|u(t)\| = 1$ .

Case (b)

Whenever  $\|a(t)\| < 1$ ,  $f(t)$  is maximized by setting  $\|u(t)\| = 0$ , or equivalently,

$$u(t) = 0$$

Since  $u(t)$  is also zero for the special case where  $\|a(t)\| = 0$ , as seen above, then  $u(t) = 0$  whenever  $0 \leq \|a(t)\| < 1$ .

Case (c)

Whenever  $\|a(t)\| = 1$ , (17) reveals that the choice of  $\|u(t)\|$  is immaterial as long as  $u(t)$  is admissible. The system is defined to be *regular* if the set of points in  $[0, \tau]$  at which  $\|a(t)\| = 1$  is of measure zero for every vector  $\eta^*$ . Only regular systems will be considered here. For such systems, the control which maximizes  $\eta^* \cdot \omega$  (for  $\omega \in \Omega(\tau)$ ) is determined almost everywhere. Since sets of measure zero play no role in the systems under consideration, they will be ignored, so that the control which maximizes  $\eta^* \cdot \omega$ , that is, the optimal control, is uniquely determined. For the sake of completeness, however, the choice  $u(t) = 0$  will be made whenever  $\|a(t)\| = 1$ .

Note that because of the uniqueness, inequality (12) can be replaced by the strict inequality

$$\eta^* \cdot y_0 > \eta^* \cdot \omega \quad (18)$$

for all  $\omega \in \Omega(\tau)$  different from  $y_0$ .

Recalling the definition of  $a(t)$  from (15), and summarizing the above results, the admissible control which maximizes  $\eta^* \cdot \omega$  in (13) is given by

$$u(t) = \begin{cases} \frac{[\eta \cdot X^{-1}(t) B(t)]'}{\|\eta \cdot X^{-1}(t) B(t)\|} & \text{if } \|\eta \cdot X^{-1}(t) B(t)\| > 1 \\ 0 & \text{if } \|\eta \cdot X^{-1}(t) B(t)\| \leq 1 \end{cases} \quad (19)$$

for  $0 \leq t \leq \tau$ .

Note from (19) that whether or not the control is 'on' is governed by whether or not  $\|\eta \cdot X^{-1}(t) B(t)\|$  exceeds a fixed threshold. Moreover, when the control is 'on', the system utilizes the full capability of this control by making  $\|u(t)\| = 1$  and changing only the 'direction' of the control in accordance with the components of  $\eta \cdot X^{-1}(t) B(t)$ .

### Synthesis of the Optimal Control

Since  $\eta$  is the only unknown in (19), the problem of synthesizing the optimal control is that of determining an  $\eta$  that corresponds to a given initial state  $x(0)$ .

Another way of looking at this problem is to observe that the transpose of the row vector  $\eta \cdot X^{-1}(t)$  is the solution of the homogeneous system adjoint to (7):

$$\dot{\psi}(t) = -A'(t) \psi(t), \quad \psi(0) = \eta'$$

where the prime denotes the transpose. Hence, the synthesis problem is equivalently that of determining the initial conditions on the system adjoint.

Now consider an  $(n+1)$ -dimensional vector of the form

$$\lambda^* = (\lambda, -1) \quad (20)$$

where  $\lambda = (\lambda_1, \dots, \lambda_n)$ .

For an arbitrary  $\lambda$ , let the function  $G(t, \lambda)$  be given by

$$G(t, \lambda) = \begin{cases} [\lambda \cdot X^{-1}(t) B(t)]' & \text{if } \|\lambda \cdot X^{-1}(t) B(t)\| > 1 \\ \|\lambda \cdot X^{-1}(t) B(t)\| & \text{if } \|\lambda \cdot X^{-1}(t) B(t)\| \leq 1 \end{cases}$$

which is obtained from (19) by replacing  $\eta$  by  $\lambda$ . For a regular system,  $G(t, \lambda)$  is unique almost everywhere for every vector  $\lambda$ .

Define the vector function

$$z(\tau, \lambda^*) = \left[ \int_0^\tau X^{-1}(t) B(t) G(t, \lambda) dt, \int_0^\tau \|G(t, \lambda)\| dt \right] \quad (21)$$

In the same way as relation (18) was derived, the relation

$$\lambda^* \cdot z(\tau, \lambda^*) > \lambda^* \cdot \xi \quad (22)$$

for all  $\xi \in \Omega(\tau)$ ,  $\xi \neq z(\tau, \lambda^*)$  can also be obtained. Hence,  $z(\tau, \lambda^*)$  is on the boundary of  $\Omega(\tau)$ . The hyperplane of support to  $\Omega(\tau)$  at  $z(\tau, \lambda^*)$  has the normal vector  $\lambda^*$ . This hyperplane is denoted by  $B(\lambda^*)$  as shown in Figure 6.

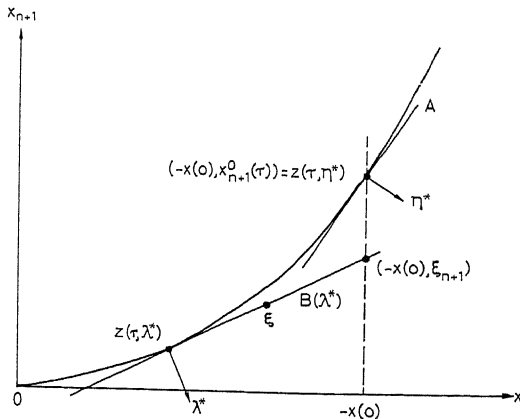


Figure 6. Representation of the problem of determining  $\eta^*$

The geometric formulation of the problem of determining an  $\eta^*$  associated with a specified initial state  $x(0)$  is depicted in Figure 6. If  $\eta^*$  is known, the corresponding point on the boundary of  $\Omega(\tau)$  has the coordinates

$$z(\tau, \eta^*) = (-x(0), x_{n+1}^0(\tau)) \quad (23)$$

where  $x_{n+1}^0(\tau)$  has been defined previously. The hyperplane of support at this point is denoted by  $A$ .

Let  $\lambda^*$  be any vector as defined in (20) such that  $z(\tau, \lambda^*) \neq z(\tau, \eta^*)$ . As shown above,  $z(\tau, \lambda^*)$  is on the boundary of  $\Omega(\tau)$ . If  $\xi = (\xi_1, \dots, \xi_n, \xi_{n+1})$  is any point in the space  $E^{n+1}$  which also lies in  $B(\lambda^*)$ , then

$$\lambda^* \cdot z(\tau, \lambda^*) = \lambda^* \cdot \xi \quad (24)$$

This is simply the equation of the plane  $B(\lambda^*)$ . The line through the point  $(-x(0), 0)$  parallel to the  $(n+1)$ st coordinate axis intersects  $B(\lambda^*)$  at the point  $\xi = (-x(0), \xi_{n+1})$  as shown in the figure. Substituting  $\xi = (-x(0), \xi_{n+1})$  into (24) gives

$$\lambda^* \cdot z(\tau, \lambda^*) = \lambda^* \cdot (-x(0), \xi_{n+1}) \quad (25)$$

Since  $z(\tau, \lambda^*) \neq z(\tau, \eta^*)$ , it follows from (22) that

$$\lambda^* \cdot z(\tau, \lambda^*) > \lambda^* \cdot z(\tau, \eta^*) \quad (26)$$

Substituting (23) and (25) into (26) gives

$$\lambda^* \cdot (-x(0), \xi_{n+1}) > \lambda^* \cdot (-x(0), x_{n+1}^0(\tau)) \quad (27)$$

Expanding (27) yields

$$-\lambda \cdot x(0) - \xi_{n+1} > -\lambda \cdot x(0) - x_{n+1}^0(\tau)$$

which simplifies to

$$\xi_{n+1} < x_{n+1}^0(\tau) \quad (28)$$

This result is geometrically clear from Figure 6 and indicates that the value of  $\xi_{n+1}$  cannot exceed  $x_{n+1}^0(\tau)$ .

Solving (25) for  $\xi_{n+1}$  gives

$$\xi_{n+1} = -\lambda^* \cdot (x(0), 0) - \lambda^* \cdot z(\tau, \lambda^*) \quad (29)$$

It is clear from (29) and Figure 6 that  $\xi_{n+1}$  is a function of  $\lambda$  since  $\lambda^* = (\lambda, -1)$ .

Now, if  $z(\tau, \lambda^*) = z(\tau, \eta^*)$ , it readily follows that

$$\xi_{n+1}(\lambda) = x_{n+1}^0(\tau) \quad (30)$$

However, (30) holds if and only if  $z(\tau, \lambda^*) = z(\tau, \eta^*)$ .

Hence, the problem of synthesizing the optimal control has been reduced to the problem of finding an  $(n+1)$ -dimensional vector  $\lambda^* = (\lambda, -1)$  which maximizes  $\xi_{n+1}$  which is given in (29).

### Computation of the Optimal Control

Since the  $(n+1)$ st coordinate of  $\lambda^*$  is fixed, the maximization of  $\xi_{n+1}$  in (29) is actually with respect to the  $n$ -dimensional vector  $\lambda$ .

From (21)

$$\lambda^* \cdot z(\tau, \lambda^*) = \int_0^\tau [\lambda \cdot X^{-1}(t) B(t) G(t, \lambda) - \|G(t, \lambda)\|] dt \quad (31)$$

in which  $\lambda$  is the only variable. To simplify the notation, let

$$g(\lambda) = \lambda^* \cdot z(\tau, \lambda^*) \quad (32)$$



Substituting (32) into (29) and expanding the result gives

$$\xi_{n+1}(\lambda) = -\lambda \cdot \tau(0) - g(\lambda) \quad (33)$$

The maximization of  $\xi_{n+1}(\lambda)$  may be performed readily by utilizing the method of steepest ascent. The variable  $\lambda$  is made a function of some parameter  $\sigma$  and the differential equation

$$\frac{d\lambda}{d\sigma} = k \nabla \xi_{n+1}(\lambda) \quad (34)$$

is solved for  $\lambda(\sigma)$ . Here  $\lambda$  is an  $n$ -dimensional column vector,  $k$  is some positive constant, and  $\nabla$  denotes the gradient. Then, under the conditions discussed below and a suitable initial choice of  $\lambda$ , that is,  $\lambda(0)$ , it can be shown that the desired  $\eta$  is given by

$$\eta = \lim_{\sigma \rightarrow \infty} \lambda(\sigma)$$

In particular, it will be shown below that if this limit does exist, it is precisely the desired  $\eta$ . Substitution of this  $\eta$  into (19) then gives the optimal control for the problem.

It can be shown<sup>8</sup> that  $\nabla \xi_{n+1}(\lambda)$  is a continuous function of  $\lambda$  and is given by

$$\nabla \xi_{n+1}(\lambda) = -x(0) - \int_0^\tau X^{-1}(t) B(t) G(t, \lambda) dt \quad (35)$$

Hence, from (34)  $d\lambda/d\sigma$  is also continuous.

Assume

$$\lim_{\sigma \rightarrow \infty} \lambda(\sigma) = \lambda^0 \quad (36)$$

where  $\lambda^0$  is a constant  $n$ -dimensional vector. Since  $\nabla \xi_{n+1}(\lambda)$  is continuous,

$$\lim_{\sigma \rightarrow \infty} \nabla \xi_{n+1}(\lambda(\sigma)) = \nabla \xi_{n+1}(\lambda^0) \quad (37)$$

From (36) and (37) it then follows that as  $\lambda \rightarrow \lambda^0$ ,  $d\lambda/d\sigma \rightarrow 0$ , i. e.

$$\lim_{\sigma \rightarrow \infty} \nabla \xi_{n+1}(\lambda(\sigma)) = 0 \quad (38)$$

Hence, for  $\lambda = \lambda^0$  in (35), it is clear that

$$-x(0) = \int_0^\tau X^{-1}(t) B(t) G(t, \lambda^0) dt \quad (39)$$

This means that  $G(t, \lambda^0)$  is precisely the control required to drive the system to the origin from  $x(0)$  in time  $\tau$ .

Let  $\lambda^{0*} = (\lambda^0, -1)$ . Then, from the definition of  $z(\tau, \lambda^*)$ , it follows that

$$z(\tau, \lambda^{0*}) = \left[ \int_0^\tau X^{-1}(t) B(t) G(t, \lambda^0) dt, \int_0^\tau \|G(t, \lambda^0)\| dt \right] \quad (40)$$

Substitution of (39) into (40) gives

$$z(\tau, \lambda^{0*}) = \left[ -x(0), \int_0^\tau \|G(t, \lambda^0)\| dt \right]$$

From (23) (see also Figure 6),

$$z(\tau, \eta^*) = (-x(0), x_{n+1}^0(\tau))$$

Hence, the first  $n$  coordinates of  $z(\tau, \lambda^{0*})$  are the same as the first  $n$  coordinates of  $z(\tau, \eta^*)$ . By hypothesis,  $z(\tau, \eta^*)$  is a

boundary point of  $\Omega(\tau)$ . From (22) it is clear that  $z(\tau, \lambda^{0*})$  is also a boundary point of  $\Omega(\tau)$ . Hence,

$$z(\tau, \lambda^{0*}) = z(\tau, \eta^*)$$

and, therefore,  $G(t, \lambda^0)$  is the optimal control.

Substituting (35) into (34) yields

$$\frac{d\lambda}{d\sigma} = -k \left[ x(0) + \int_0^\tau X^{-1}(t) B(t) G(t, \lambda) dt \right]$$

which may be rewritten as

$$\frac{d\lambda}{d\sigma} = -k X^{-1}(\tau) \left\{ X(\tau) \left[ x(0) + \int_0^\tau X^{-1}(t) B(t) G(t, \lambda) dt \right] \right\} \quad (41)$$

From (10) the term in brackets is the solution of (7) for  $x(\tau)$  where the choice of  $\lambda$  determines the control. Since this  $x(\tau)$  depends on  $\lambda$  which is a function of  $\sigma$  in (41), define

$$x(\tau, \lambda) = X(\tau) \left[ x(0) + \int_0^\tau X^{-1}(t) B(t) G(t, \lambda) dt \right] \quad (42)$$

Substituting (42) into (41) yields

$$\frac{d\lambda}{d\sigma} = -k X^{-1}(\tau) x(\tau, \lambda) \quad (43)$$

The discrete form of (43) is

$$\lambda^{(i+1)} - \lambda^{(i)} = -K X^{-1}(\tau) x(\tau, \lambda^{(i)}) \quad (44)$$

where  $K$  is some positive constant and  $i$  is the index of iteration.

A computational procedure is now clear. An initial 'guess'  $\lambda^{(0)}$  is made and substituted into (19) to determine the corresponding control. This control is then applied to the system of (7) to obtain  $x(\tau, \lambda^{(0)})$ . This result is substituted into (44) to compute  $\lambda^{(1)}$ , the first iteration on  $\lambda$ . The cycle is then repeated using (44) to iterate on  $\lambda$  each time. The process continues until  $\|\lambda^{(i+1)} - \lambda^{(i)}\|$  is less than some specified constant. Then, the optimal control is obtained from (19) using the last iteration on  $\lambda$ .

## Discussion of Results

This paper has presented a synthesis procedure for a class of minimal effort controls. While the results are fairly general, it will prove instructive to discuss them in terms of the guidance application which motivated the study.

By minimizing the fuel consumed in the midcourse phase, more space is made available on a given space vehicle for such items as life support equipment, communication systems, and scientific instruments. The scope and efficiency of space missions are, of course, directly dependent on the availability of additional equipment. While it is desirable that the space saved by using the optimal system be sizable, a saving of only a few pounds may permit the inclusion of experiments not possible otherwise. An important factor in space and weight considerations is whether or not the computations required can be done by equipment already on board the vehicle. If additional computer equipment is required, the question of savings in weight and space must be examined carefully.

Just how close the terminal error  $x(\tau)$  will come to being zero in any practical system depends on a number of factors.

Some of the more important factors include the accuracy with which the initial error  $x(0)$  can be measured, the accuracy with which the control can be computed and the capability of the midcourse propulsion system to implement this control.

The first of these factors depends upon the performance limitations of measurement schemes and the associated sensors. In general, the longer one is willing to wait, the better will be the estimates of position and velocity errors. However, in waiting, initial errors are permitted to continue uncontrolled. The trade-off is obvious. One possible approach is to divide the mission interval  $[0, \tau]$  into two or more sub-intervals. Then, while control is being executed over one sub-interval on the basis of earlier error estimates, data can be sensed and processed to improve the error estimates for succeeding sub-intervals. While the system is now optimal over each sub-interval rather than over the entire interval  $[0, \tau]$ , the control is based on more accurate position and velocity data. Since the control operates 'open-loop' once  $\lambda$  is determined, this latter approach also permits the system to control for disturbances occurring in previous sub-intervals.

The accuracy with which the control can be computed is governed by how rapidly the solution of (43) converges. In addition, the amount of computing time allowed before control must be initiated as well as computer speed are key factors. Studies are needed to determine how small  $\|\lambda^{(i+1)} - \lambda^{(i)}\|$  must be before computation is terminated. This information is essential to ensure that  $\|x(\tau)\|$  can be brought within reasonable limits.

As a result of the 'on-off' nature of the optimal control, non-throtttable propulsion can be used. Hence, the problems associated with controlling throtttable engines are completely circumvented. However, since the direction of the control is time-varying, some means must be provided for controlling this direction.

## Appendix

In this appendix it will be proved that the set  $\Omega(\tau)$  is convex, closed, and bounded. Since the components of  $X^{-1}(t)$ ,  $B(t)$ , and  $u(t)$  (so long as  $u$  is admissible) for  $0 \leq t \leq \tau$ , are uniformly bounded, it immediately follows that  $\Omega(\tau)$  is bounded. Let

$$L(u(t)) = \int_0^\tau X^{-1}(t) B(t) u(t) dt$$

and

$$F(u(t)) = \int_0^\tau \|u(t)\| dt$$

To show that  $\Omega(\tau)$  is convex it must be shown that if  $y^1$  and  $y^2$  are any two points in  $\Omega(\tau)$ , then  $y^* = \alpha y^1 + \beta y^2$  also belongs to  $\Omega(\tau)$  whenever  $\alpha \geq 0$ ,  $\beta \geq 0$ ,  $\alpha + \beta = 1$ . Since  $y^i \in \Omega(\tau)$ ,  $y^i = [L(u^i(t)), F(u^i(t))]$ , where  $u^i(t)$  is admissible ( $i = 1, 2$ ). Let  $u^*(t) = \alpha u^1(t) + \beta u^2(t)$ ; then

$$\|u^*(t)\| = \|\alpha u^1(t) + \beta u^2(t)\| \leq \alpha \|u^1(t)\| + \beta \|u^2(t)\| \leq \alpha + \beta = 1$$

for all  $t$  (since  $u^1$  and  $u^2$  are admissible), so that  $u^*(t)$  is admissible. The above inequalities also imply that  $F(u^*) \leq \alpha F(u^1) + \beta F(u^2)$ . Since  $L$  is a linear operator,  $L(u^*) = \alpha L(u^1) + \beta L(u^2)$ . Thus  $y^* = [\alpha L(u^1) + \beta L(u^2), \alpha F(u^1) + \beta F(u^2)] = [L(u^*), y^*_{n+1}]$  where  $y^*_{n+1} \geq F(u^*)$ .

An admissible control  $u^{**}(t)$  will now be constructed such that  $L(u^*) = L(u^{**})$  and  $F(u^{**}) = y^*_{n+1}$ , i.e. such that

$y^* = [L(u^{**}), F(u^{**})]$ , so that  $y^* \in \Omega(\tau)$ . Since  $\|u^i(t)\| \leq 1$  for all  $t$ , it follows that  $F(u^i) \leq \tau$  ( $i = 1, 2$ ). Thus

$$y^*_{n+1} = \alpha F(u^1) + \beta F(u^2) \leq (\alpha + \beta) \tau = \tau$$

To construct  $u^{**}$ , a control  $v(t)$  will first be constructed, with the property that  $\|v(t)\| = 1$  for all  $t$  (so that  $F(v) = \tau$ ), and such that  $L(v) = L(u^*)$ . Let

$$\bar{u}(t) = \begin{cases} u^*(t) & \text{if } u^*(t) \neq 0 \\ (1, 0, \dots, 0) & \text{if } u^*(t) = 0 \end{cases}$$

so that  $1 \geq \|\bar{u}(t)\| > 0$  for all  $t$ . It is clear that  $\bar{u}(t)$  is admissible and that

$$u^*(t) = \|u^*\| \frac{\bar{u}(t)}{\|\bar{u}(t)\|} \quad (45)$$

Let

$$\eta(t) = X^{-1}(t) B(t) \frac{\bar{u}(t)}{\|\bar{u}(t)\|} \quad (46)$$

Clearly,  $\eta(t)$  is measurable and bounded, and, by (45),

$$L(u^*) = \int_0^\tau \eta(t) \|u^*(t)\| dt$$

According to Lemma 2 of ref. 9, there is a measurable function  $\alpha(t)$  with  $|\alpha(t)| \equiv 1$  such that  $L(u^*) = \int_0^\tau \eta(t) \alpha(t) dt$ . Let

$$v(t) = \alpha(t) \frac{\bar{u}(t)}{\|\bar{u}(t)\|} \quad (47)$$

Then  $\|v(t)\| = |\alpha(t)| = 1$  for all  $t$ , and, by (46) and (47),

$$L(u^*) = \int_0^\tau \eta(t) \alpha(t) dt = \int_0^\tau X^{-1}(t) B(t) v(t) dt = L(v)$$

and  $v(t)$  has the desired properties. Repeating the above construction on sub-intervals of  $[0, \tau]$ , it is possible to find admissible controls  $v^s(t)$ , for  $0 \leq t \leq s$  ( $s < \tau$ ), such that  $\|v^s(t)\| = 1$  for all  $t \leq s$ , and

$$\int_0^s X^{-1}(t) B(t) u^*(t) dt = \int_0^s X^{-1}(t) B(t) v^s(t) dt$$

Let

$$u^s(t) = \begin{cases} v^s(t) & \text{for } 0 \leq t \leq s \\ u^*(t) & \text{for } s < t \leq \tau \end{cases}$$

Then it is easily seen that  $L(u^s) = L(u^*)$ . Define  $\theta(s) = F(u^s)$ , so that  $\theta(0) = F(u^0) = F(u^*) \leq y^*_{n+1}$  and  $\theta(\tau) = F(u^*) = \tau \geq y^*_{n+1}$ . It is obvious that  $\theta(s)$  is a continuous function of  $s$ , so that  $\theta(s_0) = y^*_{n+1}$  for some  $s_0$ ,  $0 \leq s_0 \leq \tau$ . If  $u^{**}(t) = u^{s_0}(t)$ ,  $u^{**}$  has the promised properties.

To prove that  $\Omega(\tau)$  is closed, let  $y^j = [L(u^j), F(u^j)]$ ,  $j = 1, 2, \dots$ ,  $u^j(t)$  admissible, be a sequence of points in  $\Omega(\tau)$  with  $y^j \rightarrow y^*$ . It must be shown that  $y^* \in \Omega(\tau)$ . Since  $\|u^j_i(t)\| \leq 1$  for all  $t, j$ , and  $i = 1, \dots, r$ , the functions  $u^j_i(t)$  are uniformly bounded in the norm of the Hilbert space  $L_2(0, \tau)$ . By a well-known property of  $L_2$ , there exists a subsequence  $u^{j_k}(t)$  such that  $u^{j_k}_i(t) \rightarrow u_i^*(t)$  weakly for every  $i$  and some functions  $u_i^*(t)$  in  $L_2$ . It will simply be assumed that  $u^{j_k}_i(t) \rightarrow u_i^*(t)$  weakly for each  $i$ . By definition of weak convergence,  $L(u^{j_k}) \rightarrow L(u^*)$ , where  $u^* = (u_1^*, \dots, u_r^*)$ . It is a further consequence

of the weak convergence that  $\|u^*(t)\| \leq 1$  for almost all  $t$ , and that  $F(u^*) \leq \lim_{j \rightarrow \infty} F(u^j)$ . Thus,  $y^* = [\lim L(\tilde{u}^j), \lim F(u^j)] = [L(u^*), y^*_{n+1}]$  with  $y^*_{n+1} \geq F(u^*)$ . Now an admissible control  $u^{**}(t)$  such that  $y^* = [L(u^{**}), F(u^{**})]$  can be constructed just as above, thereby proving that  $y^* \in \Omega(\tau)$  and that  $\Omega(\tau)$  is closed.

## References

- <sup>1</sup> BOLTYANSKII, V. G., GAMKRELIDZE, R. V., and PONTRYAGIN, L. S. The theory of optimal processes. 1. The maximum principle. *Amer. Math. Soc. Trans. Ser. 2*, 18 (1961), 341
- <sup>2</sup> KALMAN, R. E. The theory of optimal control and the calculus of variations. *Res. Inst. Adv. Studies Tech. rep.* 61-3
- <sup>3</sup> BERKOVITZ, L. D. Variational methods in problems of control and programming. *J. Math. Analysis and Applic.* 3, No. 1 (1961) 145
- <sup>4</sup> LUKES, D. Application of Pontryagin's maximum principle in

- determining the optimal control of a variable mass vehicle. *Amer. Rocket Soc. Paper* 1927-61, August 1961
- <sup>5</sup> LEITMANN, G. On a class of variational problem in rocket flight. *J. aero. Sci. Amer.* 26, No. 9 (1959), 586
- <sup>6</sup> ISAEV, V. K. L. S. Pontryagin's maximum principle and optimal programming of rocket thrust. *Automat. Rem. Control* 22, No. 18 (1961), 881
- <sup>7</sup> BRYSON, A. E., et al. Determination of the lift or drag program that minimizes re-entry heating with acceleration or range constraints using a steepest descent computational procedure. Paper presented at *Inst. aero. Sci. Amer. meeting*. New York; January 1961
- <sup>8</sup> NEUSTADT, L. W. On synthesizing optimal controls. *Automatic and Remote Control*, 1963. London; Butterworths: Munich; Oldenbourg
- <sup>9</sup> LA SALLE, J. P. The 'bang-bang' principle. *Automatic and Remote Control*, 1961. London; Butterworths

## DISCUSSION

M. ATHANS, M.I.T. Lincoln Laboratory, Lexington, Massachusetts, U.S.A.

In the literature on optimal control, the paper by Meditch and Neustadt is one of the few concerned with control constraints of the type  $\|u(t)\| \leq 1$ .

The purpose of this discussion is to present some recent results which provide the solution to a class of problems using the same control constraint.

Consider the system

$$\dot{x}(t) = f[x(t); t] + u(t)$$

where  $x(t)$ ,  $\dot{x}(t)$ ,  $u(t)$ , and  $f$  are  $n$  dimensional vectors and  $t$  is the (scalar) time.

**Assumption 1** The control vector  $u(t)$  has  $n$  non-zero components  $u_1(t), \dots, u_n(t)$  and is constrained by

$$\|u(t)\| = \sqrt{u_1^2(t) + \dots + u_n^2(t)} \leq M \text{ for all } t$$

**Assumption 2**

$$\langle f[x(t); t], x(t) \rangle = g[\|x(t)\|; t]$$

for all  $x(t)$  and  $t$ , where  $\langle f, x \rangle$  is the scalar product of the vectors  $f$  and  $x$ ,  $\|x(t)\|$  is the Euclidean norm of the vector  $x(t)$  and  $g$  is the scalar function of  $\|x(t)\|$  and  $t$ .

Under these assumptions it has been proved<sup>1</sup> that the control

$$u(t) = -M \frac{x(t)}{\|x(t)\|}$$

will force any initial controllable state to the origin, 0, in minimum time.

In the special case

$$g[\|x(t)\|; t] = 0$$

for all  $x(t)$  and  $t$ , the time-optimal control  $u(t) = -Mx(t)/\|x(t)\|$  is also fuel-optimal to the origin in the sense that it minimizes the functional

$$\int_0^T \|u(t)\| dt$$

This fuel-optimality has also been proved<sup>2, 3</sup>.

The theoretical results given above can be used for the optimal angular velocity control of a tumbling body in space, using gimbaled reaction jets.

## References

- <sup>1</sup> ATHANS, M. and FALB, P. L. Time-optimal control for a class of non-linear systems. *I.E.E.E. Trans. Automatic Control*. (October 1963)
- <sup>2</sup> ATHANS, M., FALB, P. L. and LACOSS, R. T. Time-, fuel-, and energy-optimal control of non-linear norm invariant systems. *I.E.E.E. Trans. Automatic Control*. (July 1963)
- <sup>3</sup> ATHANS, M., FALB, P. L. and LACOSS, R. T. On optimal control of self-adjoint systems. *Proc. 1963 Joint Automatic Control Conf.* (June 1963), pp. 113-120

J. S. MEDITCH, in reply

Dr. Athans' interesting discussion is especially important since it presents one of the few classes of problems in which the minimal fuel control law can be obtained in closed form as a function of the 'instantaneous' state of the dynamical system, i.e., as a feedback law.

# On Necessary and Sufficient Conditions for Time Optimal Control of Second-order Non-linear Systems

E. B. LEE and L. MARKUS

## Summary

Problems of constructing optimal control for non-linear second-order systems are considered. It is shown that, under appropriate hypotheses, extremal control for steering from a given initial point to a target is unique. This result is then used with previously known results and geometric methods to establish the global uniqueness of extremal control for certain non-linear second-order systems. It is shown how the optimum feedback control function can be constructed for this class of systems. Questions of existence and structural properties of optimal control are treated.

## Sommaire

Les problèmes de la réalisation d'un réglage optimal pour les systèmes non-linéaires du deuxième ordre sont examinés. Les résultats prouvent qu'à partir d'une hypothèse convenable, le réglage extrême pour guidage à partir d'un point donné jusqu'au but, est unique. Ce résultat est employé, avec des résultats antérieurement connus et avec des méthodes géométriques, pour établir l'unicité globale du réglage extrême pour certains systèmes non-linéaires du deuxième ordre. La réalisation de la fonction optimale de réaction pour des systèmes de cette classe est ensuite développée. Sont aussi discutées les questions d'existence et de propriétés de structure du réglage optimal.

## Zusammenfassung

Der Aufsatz behandelt die Probleme der optimalen Regelung von nichtlinearen Systemen zweiter Ordnung. Es zeigt sich, daß unter geeigneten Annahmen ein eindeutiger optimaler Regelvorgang existiert, um von einem gegebenen Ausgangspunkt zu einem Ziel zu gelangen. Dieses Ergebnis wird zusammen mit bereits bekannten Resultaten und geometrischen Methoden dazu benutzt, Extremwertregelungen bestimmter nichtlinearer Systeme 2. Ordnung mit umfassender Eindeutigkeit zu erstellen.

Die Arbeit zeigt, wie der optimale Regelkreis für diese Klasse von Systemen aussehen muß. Fragen der Existenz und strukturelle Eigenschaften der optimalen Regelung werden behandelt.

## Introduction

There are many physical systems whose evolution can be summarized as a solution of an ordinary second order differential equation. In certain of these systems there are parameters, due to design or otherwise, which can be selected to change the evolution of the system or analogously to change a solution of the differential equation. These control parameters may be subject to certain restrictions, for example, in the case to be considered they are bounded in amplitude and assumed measurable. Interest here is in constructing a controller (in terms of the differential equation, a function of two variables) which determines the control parameter so that each solution of the second-order differential equation passes to the origin (rest position) in finite time, in fact, the shortest possible time.

In the next section the problem is formulated and the necessary conditions for time optimal control indicated. The optimal controls are shown to be 'bang-bang' with switches in magnitude occurring in a cyclic order with the zeros of the velocity coordinate when viewed as a function of time. There are certain questions concerning the existence of optimal control. Two important cases of the question of existence are resolved.

The construction of a switching locus in the phase plane is considered as a possible way of constructing the optimal controller. It is shown that the switching locus contains a piecewise smooth curve which separates the phase plane. Cases in which the construction of the switching locus based on the necessary condition is also sufficient for optimal control are treated. An example is also shown.

These results are essentially those which have been presented in detail<sup>6, 7</sup>. The purpose here is to summarize the results for this problem and to explain their use and physical significance. Most of the proofs can be found in Lee and Markus<sup>6, 8</sup>.

## Problem Statement and the Necessary Conditions

Consider a dynamical system as described by the real second order differential equation

$$\ddot{x} + f(x, \dot{x}) = u$$

or the equivalent system

$$\dot{x} = y$$

$$(S) \quad \dot{y} = -f(x, y) + u$$

Here  $u$  is a scalar control variable, which is subject to the restriction  $-1 \leq u \leq 1$ , and the real function  $f(x, y)$  is assumed differentiable everywhere with respect to the two variables  $x$  and  $y$ . The  $(x, y)$  plane is denoted here by  $R^2$ , the two-dimensional real number space.

It is desired to synthesize  $\Phi(x, \dot{x})$  in the differential equation

$$\ddot{x} + f(x, \dot{x}) = \Phi(x, \dot{x})$$

so that each of its responses  $(x(t), \dot{x}(t)) = (x(t), y(t)) \rightarrow (0, 0)$  in the minimum time. Generally the authors find controls  $u = u(t)$  for the system (S) so that the response  $(x(t), y(t)) \rightarrow (0, 0)$  in minimum time under the influence of  $u(t)$  and then indicate how  $u(t)$  can be used to construct the feedback controller  $\Phi(x, y)$ .

Consider the dynamical system (S). For a given initial condition  $(x_0, y_0)$  at time  $t = 0$ , define  $\Delta$  as the class of all (measurable) controls  $u(t)$  on various time intervals  $0 \leq t \leq t_1$ , with  $-1 \leq u(t) \leq 1$ , for each  $0 \leq t \leq t_1$  such that the response  $\bar{x}(t) = (x(t), y(t))$  is defined on  $0 \leq t \leq t_1$  with  $x(0) = x_0$ ,  $y(0) = y_0$ , and first reaches the origin at  $t = t_1$ . The response  $(x(t), y(t))$  is an absolutely continuous solution of (S) with  $x(0) = x_0$ ,  $y(0) = y_0$ .

A control  $u(t)$  on  $0 \leq t \leq t_1$  in  $\Delta$  is called (minimal) time optimal in case, for each  $\hat{u}(t)$  in  $\Delta$  it is found that  $t_1 \leq \hat{t}$ .

By the Maximum Principle<sup>1</sup> an optimal control is necessarily an extremal control. That is:

(i) There exists a nowhere zero, absolutely continuous, response  $\bar{\eta}(t) = (\eta_1(t), \eta_2(t))$  on  $0 \leq t \leq t_1$  such that  $\bar{x}(t)$ ,  $\bar{\eta}(t)$ ,  $u(t)$  satisfy the equation system

$$\begin{cases} \dot{x} = \frac{\partial H}{\partial \eta_1} \\ \dot{y} = \frac{\partial H}{\partial \eta_2} \end{cases} \quad \begin{cases} \dot{\eta}_1 = -\frac{\partial H}{\partial x} \\ \dot{\eta}_2 = -\frac{\partial H}{\partial y} \end{cases}$$

and,

(ii)  $H(\bar{\eta}(t), \bar{x}(t), u(t)) = M(\eta(t), \bar{x}(t))$  for almost all  $t$  on  $0 \leq t \leq t_1$ . Here  $H(\bar{\eta}, \bar{x}, u) = \eta_1 \dot{\bar{x}} + \eta_2 \dot{\bar{y}} = \eta_1 y + \eta_2 [-f(x, y) + u]$  and  $M(\bar{\eta}, \bar{x}) = \max_{-1 \leq u \leq 1} H(\bar{\eta}, \bar{x}, u)$ .

Moreover,

(iii)  $M(\bar{\eta}(0), \bar{x}(0)) = M(\bar{\eta}(t), \bar{x}(t)) \geq 0$  for all  $t$  on  $0 \leq t \leq t_1$ , and  $\bar{\eta}(t)$ ,  $\bar{x}(t)$  satisfying (i) and (ii).

The authors state that a control  $u(t)$  on  $0 \leq t \leq t_1$  in  $\Delta$  is a relay control in case there exists a finite number of times  $0 = \tau_0 \leq \tau_1 \leq \dots \leq \tau_k = t_1$ , such that on each open internal  $\tau_{i-1} < t < \tau_i$ ,  $i = 1, \dots, k$ ,  $u(t)$  equals either  $+1$  or  $-1$  and switches value on successive intervals.

From the Maximum Principle the following result is obtained.

**Corollary**—Let  $u(t)$  on  $0 \leq t \leq t_1$  in  $\Delta$  be an extremal control for the system (S). Then  $u(t)$  is equal to the relay control  $\text{sgn}\{\eta_2(t)\}$  almost everywhere on  $0 \leq t \leq t_1$ . (Hereafter it is to be assumed that the extremal control  $u(t)$  has been modified on a null set so as to be equal everywhere to the corresponding relay control; that is,  $u(t) = \text{sgn}\{\eta_2(t)\}$ . Here  $\text{sgn}[\cdot] = 1$  if  $[\cdot] > 0$ ,  $-1$  if  $[\cdot] < 0$ ,  $0$  if  $[\cdot] = 0$ .) Also the response  $\eta_2(t)$  on  $0 \leq t \leq t_1$  has a finite number of zeros each of which is simple.

From the Maximum Principle and certain results from the Sturm oscillation theory, the switching of the relay control can also be related to the geometry of the  $(x, y)$  plane. This result, which states that the zeros of  $y(t)$  and  $\eta_2(t)$  are interlaced, is basic in establishing the properties of the switching locus as dealt with later.

**Theorem 1**—Let  $u(t)$  on  $0 \leq t \leq t_1$  in  $\Delta$  be an extremal control for the system (S), and let  $\bar{x}(t) = (x(t), y(t))$  and  $\bar{\eta}(t) = (\eta_1(t), \eta_2(t))$  be the corresponding extremal responses. Let  $\xi_1, \xi_2$  be times such that  $0 \leq \xi_1 < \xi_2 \leq t_1$ . Then

- (i) if  $\eta_2(\xi_1) = \eta_2(\xi_2) = 0$  and if  $y(\xi_1) = 0$ , then  $y(\xi_2) = 0$ .
- (ii) If  $\eta_2(\xi_1) = \eta_2(\xi_2) = 0$  and if  $y(\xi_1) \neq 0$ , then  $y(\xi_2) \neq 0$  but there is a zero of  $y(t)$  on the open interval  $\xi_1 < t < \xi_2$ .
- (iii) If  $y(\xi_1) = y(\xi_2) = 0$ ,  $y(t) \neq 0$  on  $\xi_1 < t < \xi_2$  and if  $\eta_2(\xi_1) = 0$ , then  $\eta_2(\xi_2) = 0$ .
- (iv) If  $y(\xi_1) = y(\xi_2) = 0$ ,  $y(t) \neq 0$  on  $\xi_1 < t < \xi_2$  and if  $\eta_2(\xi_1) \neq 0$ , then  $\eta_2(\xi_2) \neq 0$ , but there is a zero of  $\eta_2(t)$  on the open interval  $\xi_1 < t < \xi_2$ .

Thus if the zeros of  $y(t)$  are isolated, they either coincide with the zeros of  $\eta_2(t)$  or else no zero of  $y(t)$  is a zero of  $\eta_2(t)$  but these two sets of zeros are interlaced.

**Corollary**—Let  $u(t)$  on  $0 \leq t \leq t_1$  in  $\Delta$  be an extremal control for the system (S). Let the corresponding response be

$\bar{x}(t) = (x(t), y(t))$  and adjoint response be  $\bar{\eta}(t) = (\eta_1(t), \eta_2(t))$  on  $0 \leq t \leq t_1$ . Let  $\eta_2(\xi) = 0$  at  $t = \xi$ ,  $0 \leq \xi \leq t_1$ .

If  $y(\xi) > 0$ , then  $\dot{\eta}_2(\xi) < 0$ .

If  $y(\xi) < 0$ , then  $\dot{\eta}_2(\xi) > 0$ .

**Remark**—Already the optimal controls have been limited to being relay controls which switch magnitude once and only once in the upper half (lower half) plane along any optimal response which enters and eventually leaves the upper half (lower half) plane. The switches, if they occur, are from  $+1$  to  $-1$  in the upper half plane and from  $-1$  to  $+1$  in the lower half plane for  $t$  increasing. Thus the optimal responses are composed of segments of the solution curves of

$$\begin{aligned} \dot{x} &= y \\ (S_{\pm}) \quad \dot{y} &= -f(x, y) \pm 1 \end{aligned}$$

with switches determined by the switching locus  $W$  which is considered later. Hereafter  $S_+$  refers to the system (S) with  $u \equiv +1$  and  $S_-$  when  $u \equiv -1$ .

## Domain of Controllability and the Existence of Optimal Control

Consider the above system (S). The set  $\mathcal{C}$  of all points  $(x_0, y_0)$  in  $R^2$  for which there exists a measurable control  $-1 \leq u(t) \leq 1$  on a finite interval,  $0 \leq t \leq t_1$ , steering  $(x(t), y(t))$  from  $(x_0, y_0)$  to the origin is called the domain of (null) controllability. Here 'steering' means that there exists a control  $u(t)$  on  $0 \leq t \leq t_1$ , which, when substituted into the equation system (S) leads to a response  $(x(t), y(t))$  that starts at  $x(0) = x_0$ ,  $y(0) = y_0$  and passes through the point  $(0, 0)$  at  $t = t_1$ .

**Theorem 2**—For the system (S) as above with  $f(0, 0) = 0$ , and  $-1 \leq u \leq 1$ , the domain of controllability  $\mathcal{C}$  is an open connected subset of  $R^2$  containing the origin.

**Theorem 3**—Consider the system (S) with  $-1 \leq u(t) \leq 1$  as above. Assume that  $f(x, y)$  is an attractive force with non-negative friction, that is

$$xf(x, 0) > 0 \quad \text{for } x \neq 0$$

and

$$\frac{\partial f}{\partial y}(x, y) \geq 0 \quad \text{in } R^2.$$

Then the domain of controllability  $\mathcal{C} = R^2$  and, moreover, each point of  $R^2$  can be steered to the origin by an optimal control  $u(t)$  in  $\Delta$ .

For the study of systems with a repulsive force the following conditions are assumed, which could be relaxed slightly in the manner suggested by Theorem 3.

$$\begin{aligned} (S_0) \quad \dot{x} &= y \\ \dot{y} &= -f(x, y), f(0, 0) = 0, f(x, y) \end{aligned}$$

differentiable and  $\partial f / \partial x(x, 0) < -\epsilon < 0$ ,  $\partial f / \partial y \geq 0$  for some  $\epsilon > 0$ .

This system has the origin as the unique critical point. Near the origin the solution curve family of  $(S_0)$  is topologically equivalent<sup>8</sup> to the solution curve family of the linear system

$$\begin{aligned} (L) \quad \dot{x} &= y \\ \dot{y} &= x - y \end{aligned}$$

Thus there are four solutions of  $(S_0)$  which approach the origin as  $t \rightarrow \infty$  or  $t \rightarrow -\infty$ . Call these curves I, II, III, and IV, depending on the quadrant in which they lie. A study of the geometry of  $(S_0)$ , cf. reference 8, shows that I and III are defined and single valued over a half  $x$  axis. Also II and IV are single valued over an interval of the  $x$  axis and on these  $|y| \rightarrow \infty$  as  $|x|$  increases.

As in their previous work<sup>7, 8</sup>, the authors find that the critical point and the curves I, II, III, IV are the only separatrices of  $(S_0)$  and the remaining solution curves of  $(S_0)$  consist of four parallel canonical regions in the sense of Markus<sup>9</sup>. Thus  $(S_0)$  is globally homeomorphic to  $(L)$  in  $R^2$ . II and IV are called the principal separatrices of  $(S_0)$ .

**Theorem 4**—Consider the system

$$(S) \quad \begin{aligned} \dot{x} &= y \\ \dot{y} &= -f(x, y) + u, f(0, 0) = 0, f(x, y) \end{aligned}$$

differentiable and  $-1 \leq u \leq 1$ . Assume  $f(x, y)$  is a repulsive force with non-negative friction, that is,

$$\frac{\partial f}{\partial x}(x, 0) < -\varepsilon < 0, \frac{\partial f}{\partial y}(x, y) \geq 0 \text{ in } R^2 \text{ for some } \varepsilon > 0$$

Then each of  $(S_+)$  and  $(S_-)$  is globally homeomorphic with the linear system  $(L)$  and has corresponding separatrices  $I_+, II_+, III_+, IV_+$  and  $I_-, II_-, III_-, IV_-$ , as described above. The domain of controllability  $\mathcal{C}$  is precisely the open topological band  $\mathcal{B}$  bounded by the principal separatrices  $II_+, IV_+$  and  $II_-, IV_-$  and the two critical points of  $S_+$  and  $S_-$ . Furthermore, each point of  $\mathcal{B}$  can be steered to the origin by an optimal control  $u(t)$  in  $\mathcal{A}$ .

### Construction and Properties of the Switching Locus

Consider the set of all extremal relay controls in  $\mathcal{A}$ , for the system  $(S)$ , steering points of the domain  $\mathcal{C}$  of controllability to the origin as above. The switching locus  $W$  is the set of all points in  $\mathcal{C}$  at which the corresponding responses  $\bar{x}(t)$  fail to have derivatives, that is,  $W$  consists of the points at which the extremal responses switch from the solution family of  $S_+$  to  $S_-$ , or vice versa.

For definiteness the origin in  $W$  is included. A construction for  $W$  is now described. Let  $W_+^1$  be the solution (or segment of a solution) of  $S_+$  through the origin and lying in the fourth quadrant,  $x \geq 0, y \leq 0$ . Let  $W_-^1$  be the solution (or segment of a solution) of  $S_-$  through the origin and lying in the second quadrant  $x \leq 0, y \geq 0$ .

Now reflect  $W_+^1$  through a conjugate interval along solutions of  $S_-$ . That is, from each point on  $W_+^1$  follow backwards in time along the corresponding solution of  $S_-$  for a time duration equal to the interval between zeros of  $\eta_2(t)$ . This means that

$$\dot{\eta}_1 = \eta_2 \frac{\partial f}{\partial x}(x(t), y(t))$$

$$\dot{\eta}_2 = -\eta_1 + \eta_2 \frac{\partial f}{\partial y}(x(t), y(t))$$

is used with  $\bar{x}(t) = (x(t), y(t))$  the appropriate solution of  $S_-$  and  $\eta_1(0) = -1, \eta_2(0) = 0$  (note that the adjoint linear system is homogeneous and that the response  $\bar{x}(t)$  to the autonomous

system  $S_-$  is started at  $t = 0$ ). The first zero of  $\eta_2(t)$  in  $t < 0$  is obtained and this determines the time interval through which one follows the solution  $\bar{x}(t)$  of  $S_-$  which initiates on  $W_+^1$ . Denote the endpoints of such solutions of  $S_-$  as the reflection  $W_-^2$  of  $W_+^1$  (see Figure 1).

Now reflect  $W_-^1$  through a conjugate interval along solutions of  $S_+$ . That is, from each point on  $W_-^1$  follow backwards in time along the corresponding solution of  $S_+$  for a time duration equal to the interval between zeros of  $\eta_2(t)$ . This means that

$$\dot{\eta}_1 = \eta_2 \frac{\partial f}{\partial x}(x(t), y(t))$$

$$\dot{\eta}_2 = -\eta_1 + \eta_2 \frac{\partial f}{\partial y}(x(t), y(t))$$

is used with  $\bar{x}(t) = (x(t), y(t))$  the appropriate solution of  $S_+$  initiating on  $W_-^1$  and  $\eta_1(0) = 1, \eta_2(0) = 0$ , to obtain the first zero of  $\eta_2(t)$  in  $t < 0$ . Denote the reflection of  $W_-^1$  along  $S_+$  by  $W_+^2$ .

If  $W_+^1, W_+^2, \dots, W_+^k$  and  $W_-^1, W_-^2, \dots, W_-^k$  have been defined, let  $W_+^{k+1}$  be the reflection of  $W_-^k$  by the solutions of  $S_+$  and let  $W_-^{k+1}$  be the reflection of  $W_+^k$  by solutions of  $S_-$ , through conjugate time intervals as indicated above. Of course it might happen that  $W_+^k \cup W_-^k$  is empty for every  $k$  greater than some positive integer.

**Theorem 5**—Consider the system  $(C^1 \text{ or } C^2 \dots \text{ indicating the degree of differentiability})$

$$(S) \quad \begin{aligned} \dot{x} &= y \\ \dot{y} &= -f(x, y) + u, f(0, 0) = 0, f(x, y) \text{ in } C^1 \end{aligned}$$

with measurable control  $u(t)$  in  $-1 \leq u \leq 1$ . The switching locus  $W$  is precisely the union of the sets  $W_+^k \cup W_-^k$  for  $k = 1, 2, 3, \dots$ , as described above.

**Proof**—Using Theorem 1 it is noted that for each point  $P$  on  $W_+^1$  in  $y < 0$  can be selected an adjoint response  $\bar{\eta}(t) = (\eta_1(t), \eta_2(t))$  with  $\eta_2(t)$  vanishing when  $\bar{x}(t) = P$  and nowhere else on  $W_+^1$  in  $y < 0$ . Also if  $P$  is an endpoint of  $W_+^1$  where  $y = 0$ , then  $\bar{\eta}(t)$  can be selected so that  $\eta_2(t)$  vanishes at times corresponding to  $\bar{x}(t) = P$  and  $\bar{x}(t) = 0$ . Then each point of  $W_+^1$ , including the endpoints, can occur as a switching point for a response to an extremal relay controller in  $\mathcal{A}$ . Thus  $W_+^1 \subset W$  and similarly  $W_-^1 \subset W$ .

Since every extremal relay control which steers a point of  $\mathcal{C}$  to the origin must have a response which enters the origin along either  $W_+^1$  or  $W_-^1$ , and since the response must follow alternately the solution curve families of  $S_+$  and  $S_-$ , or vice versa, with switches occurring at the zeros of  $\eta_2(t)$ , it is concluded that  $W$  is the union of all  $W_+^k$  and  $W_-^k$  for  $k = 1, 2, 3, \dots$  (see Figure 1).

The smoothness properties of the switching locus will now be discussed. First consider the system

$$(S) \quad \begin{aligned} \dot{x} &= y \\ \dot{y} &= -f(x, y) + u, f(0, 0) = 0, f(x, y) \text{ in } C^1 \end{aligned}$$

with measurable controllers  $u(t)$  on  $-1 \leq u \leq 1$ , and

$$\frac{\partial f}{\partial x}(x, 0) > \varepsilon > 0, \frac{\partial f}{\partial y}(x, y) \geq 0$$

for some  $\varepsilon > 0$ . These hypotheses, somewhat stronger than those of Theorem 3, assure that each of the extremal systems  $S_+$  and  $S_-$  has exactly one critical point, and these hypotheses generally simplify the exposition of the problem. For example, every solution curve of  $S_+$ , other than the unique critical point, is either a periodic solution, or spirals inwards towards a limit cycle or approaches the critical points as  $t \rightarrow +\infty$ . If, further,  $\partial f/\partial y > 0$ , then the area condition of Bendixson states that there are no limit cycles.

**Theorem 6**—Consider the system

$$(S) \quad \begin{aligned} \dot{x} &= y \\ \dot{y} &= -f(x, y) + u, f(0, 0) = 0, f(x, y) \text{ in } C^2 \end{aligned}$$

with  $\partial f/\partial x(x, 0) > \varepsilon > 0$ , for some  $\varepsilon > 0$ , and  $\partial f/\partial y \geq 0$  in  $R^2$ . Consider extremal controllers  $u(t)$  in  $\Delta$  steering points of  $\mathcal{C} = R^2$  to the origin and construct the switching locus  $W$  as the union of the sets  $W_+^k \cup W_-^k$ , as described in Theorem 5. Then  $W$  contains a homeomorph of a line, a piecewise  $C^1$  curve which separates the plane. Also  $W$  lies in  $y \leq 0$  for  $x \geq 0$  and in  $y \geq 0$  for  $x \leq 0$ .

Secondly, consider the control problem for the repulsive force case

$$(S) \quad \begin{aligned} \dot{x} &= y \\ \dot{y} &= -f(x, y) + u, f(0, 0) = 0, f(x, y) \text{ in } C^1 \end{aligned}$$

and  $u(t)$  measurable on  $-1 \leq u \leq 1$ , as above. Assume  $\partial f/\partial x(x, y) < -\varepsilon < 0$ ,  $\partial f/\partial y(x, y) \geq 0$  in  $R^2$ , for some  $\varepsilon > 0$ . These hypotheses, somewhat stronger than those of Theorem 4, assure the existence of a band  $\mathcal{B}$  between the principal separatrices of  $S_+$  and  $S_-$ , which is the domain  $\mathcal{C}$  of controllability. For each point  $P \in \mathcal{B}$  there exists an optimal control  $u(t)$  in  $\Delta$  which steers  $P$  to the origin.

The switching locus  $W$  of  $S$  is defined as above. To construct  $W$  let  $W_+^1$  be the solution of  $S_+$  through the origin and lying in the fourth quadrant. By a comparison of the slopes of the solutions of  $S_+$  and  $S_-$ , it is found that  $W_+^1$  lies always interior to the band  $\mathcal{B}$  in  $y < 0$  (except for the origin) and  $y \rightarrow \infty$  as  $x$  increases on  $W_+^1$ . Let  $W_-^1$  be the solution of  $S_-$  through the origin and lying in the second quadrant. Clearly  $W_-^1$  lies interior to  $\mathcal{B}$  and  $y \rightarrow +\infty$  as  $|x|$  increases on  $W_-^1$ . The reflections  $W_+^k$  are defined as above but it can easily be shown that these are all empty for  $k = 2, 3, 4, \dots$  and that  $W = W_+^1 \cup W_-^1$ .

**Theorem 7**—Consider the system

$$(S) \quad \begin{aligned} \dot{x} &= y \\ \dot{y} &= -f(x, y) + u, f(0, 0) = 0, f(x, y) \text{ in } C^1 \end{aligned}$$

with  $\partial f/\partial x(x, y) < -\varepsilon < 0$ ,  $\partial f/\partial y(x, y) \geq 0$  in  $R^2$ , for some  $\varepsilon > 0$ . Consider extremal relay controls  $u(t)$  on  $-1 \leq u \leq 1$  in  $\Delta$  and construct the switching locus  $W$ . Then  $W = W(x)$  is a continuous single-valued curve over a segment of the  $x$  axis, and  $W$  separates  $\mathcal{B}$  in two regions. Moreover

$$W(x) = \begin{cases} W_+^1(x) & \text{for } x \geq 0 \\ W_-^1(x) & \text{for } x \leq 0 \end{cases}$$

so  $W(x)$  in  $C^1$  except at  $x = 0$ .

## Uniqueness of Extremal Control (The Sufficient Conditions)

**Corollary to Theorem 6**—Consider the system  $S$  of Theorem 6. Assume that  $W = W(x)$  is a single-valued curve on  $-\infty < x < \infty$ . Then for each point  $P$  in  $R^2$  there is one and only one extremal relay control in  $\Delta$  which steers  $P$  to the origin.

Using this corollary it is easy to verify the following theorem.

**Theorem 8**—Consider the system

$$(S) \quad \begin{aligned} \dot{x} &= y \\ \dot{y} &= -f(x, y) + u, f(0, 0) = 0, f(x, y) \text{ in } C^1 \end{aligned}$$

with  $\partial f/\partial x(x, 0) > \varepsilon > 0$ , for some  $\varepsilon > 0$ , and  $\partial f/\partial y \geq 0$  in  $R^2$ . Assume the switching locus  $W = W(x)$  is single valued on  $-\infty < x < \infty$ . Then for each point  $P$  in  $R^2$  there is exactly one optimal control  $u(t)$  on  $-1 \leq u \leq 1$  in  $\Delta$  which steers  $P$  to the origin. (The same conclusion holds if the switching locus  $W$  is single valued with respect to both flows  $S_+$  and  $S_-$ , except for the initial curves  $W_+^1$ , and  $W_-^1$ .) Define the synthesizer

$$\Phi(x, y) = \begin{cases} 1 & \text{for } y < W(x) \\ 0 & \text{for } y = W(x) \\ -1 & \text{for } y > W(x) \end{cases}$$

Then the optimal response for  $P$  is the unique solution of

$$\ddot{x} + f(x, \dot{x}) = \Phi(x, \dot{x})$$

initiating at  $P$  at  $t = 0$ .

**Special Hypothesis**—Here it has been assumed that the switching locus  $W = W(x)$  is a single-valued curve for  $-\infty < x < \infty$ . This certainly holds if  $W_+^1$  and  $W_-^1$  extend to infinity without re-intersecting the  $x$  axis, and the adjoint equations are disconjugate, as in the next theorem, or in special cases where the damping is linear. The hypothesis also holds in the case of isochronous periodic solutions for  $S_\pm$ , for example  $f(x, y) = x$ . For here consider the variational equations along  $S_+$  or  $S_-$ ;

$$\dot{v}^1 = v^2$$

$$\dot{v}^2 = -\frac{\partial f}{\partial x} v^1 - \frac{\partial f}{\partial y} v^2$$

where  $\bar{v}(t) = (v^1(t), v^2(t))$  represents a vector and its parallel transport along the flow of  $S_+$ . But

$$v^1 \eta_1 + v^2 \eta^2 = \text{constant}$$

and so

$$v^1(0) \eta_1(0) = v^1(t_1) \eta_1(t_1)$$

for successive switching times  $t = 0$  and  $t = t_1$ . Now

$$\eta_1(0) \cdot \eta_1(t_1) < 0$$

and so

$$v^1(0) \cdot v^1(t_1) \leq 0.$$

Therefore, taking  $\bar{v}(0)$  as the tangent vector along  $W_\pm^k$ ,  $\bar{v}(t_1)$  is the tangent vector along  $W_\pm^{k+1}$  and thus  $W_\pm^{k+1}$  is single-valued over the  $x$ -axis and  $W_\pm^k$  is single-valued over the  $x$ -axis. By induction we can then show that  $W = W(x)$  is single-valued on  $-\infty < x < \infty$ .

*Corollary to Theorem 7*—Consider the system  $S$  of Theorem 7. Then for each point  $P$  in  $\mathcal{B}$  there is one and only one extremal relay control in  $\Delta$  which steers  $P$  to the origin.

*Theorem 9*—Consider the system

$$(S) \quad \begin{aligned} \dot{x} &= y \\ \dot{y} &= -f(x, y) + u, f(0, 0) = 0, f(x, y) \text{ in } C^1 \end{aligned}$$

with  $\partial f / \partial x(x, y) < -\varepsilon < 0$ , for some  $\varepsilon > 0$ , and  $\partial f / \partial y(x, y) \geq 0$  in  $R^2$ . Consider the domain  $\mathcal{C} = \mathcal{B}$  of controllability for measurable controllers  $u(t)$  on  $-1 \leq u \leq 1$  in  $\Delta$ , and construct the switching locus  $W = W(x)$ .

Then for each point  $P$  in  $\mathcal{B}$  there exists exactly one optimal control  $u(t)$  in  $\Delta$  which steers  $P$  to the origin. Define the synthesizer

$$\Phi(x, y) = \begin{cases} 1 & \text{for } y < W(x) \\ 0 & \text{for } y = W(x) \\ -1 & \text{for } y > W(x) \end{cases}$$

Then the optimal response for  $P$  is the unique solution of

$$\ddot{x} + f(x, \dot{x}) = \Phi(x, \dot{x})$$

initiating at  $P$  at  $t = 0$ .

*Remarks*—The switching locus is single valued in the case in which the damping enters linearly and is sufficiently large. Consider therefore

$$(S) \quad \begin{aligned} \dot{x} &= y \\ \dot{y} &= -g(x) - by + u(t), -1 \leq u(t) \leq 1 \end{aligned}$$

where  $b > 0$  is constant and  $g(x)$  in  $C^1$  with  $g(0) = 0$ ,  $g'(x) > 0$  and  $|g(x)| > 1$  for large  $|x|$ .

The domain  $\mathcal{C}$  of controllability is an open, connected subset of  $R^2$  which contains the origin. More precisely  $\mathcal{C} = R^2$  as in Theorems 2 and 3 above. Furthermore, each point in  $R^2$  can be steered to the origin by an optimal control  $u(t)$ , which can be assumed to be a relay control;  $u(t) = \text{sgn } \eta_2(t)$ . Thus the extremal systems are considered

$$(S_{\pm}) \quad \begin{aligned} \dot{x} &= y \\ \dot{y} &= -g(x) - by \pm 1 \end{aligned}$$

Each of the systems  $(S_{\pm})$  has the corresponding unique critical point  $0_{\pm} = (x_{\pm}, 0)$ , where  $x_{+} > 0$ ,  $x_{-} < 0$ . Also every solution curve tends towards the critical point as  $t \rightarrow +\infty$ .

Consider the solution curve  $W_{+}^1$  of  $S_{+}$  passing through  $(0, 0)$ . If  $W_{+}^1$  re-intersected the  $x$  axis, then the variational equation based on  $W_{+}^1$  has a solution  $\dot{x}(t)$  which vanishes twice. Thus consider the variational equation

$$\ddot{v} + b\dot{v} + g'(x(t))v = 0$$

Let  $z = e^{bt/2} v$  and compute

$$\ddot{z} + \left( g'(x(t)) - \frac{b^2}{4} \right) z = 0$$

Now assume  $g'(x) \leq b^2/4$  everywhere, then  $z(t)$  and hence  $v(t)$  has at most one zero. Thus  $W_{+}^1$  does not re-intersect the  $x$  axis and an easy estimate of  $dy/dx$  shows that  $W_{+}^1$  is defined and is single valued for  $x > 0$ . A similar result holds for  $W_{-}^1$  over the axis  $x < 0$ .

If the solutions  $\eta_2(t)$  of the adjoint equations can have at most one zero, then the switching locus  $W = W(x)$  is exactly  $W_{+}^1 \cup W_{-}^1$ . Consider the adjoint system

$$\dot{\eta}_1 = \eta_2 g'(x(t))$$

$$\dot{\eta}_2 = -\eta_1 + b\eta_2$$

where  $x(t)$  is some solution of  $S_{+}$  or  $S_{-}$ . Then

$$\ddot{\eta}_2 - b\dot{\eta}_2 + g'(x(t))\eta_2 = 0$$

But the solutions  $\eta_2(t)$  have the same zeros as the solutions  $z = e^{-bt/2} \eta_2$  of

$$\ddot{z} + \left( g'(x(t)) - \frac{b^2}{4} \right) z = 0$$

Thus no (non-trivial) solution  $\eta_2(t)$  has two zeros. Hence

$$W = W(x) \text{ is exactly } W_{+}^1 \cup W_{-}^1$$

*Example*— $\ddot{x} + 2\dot{x} + \arctan x = u(t)$ ,  $-1 \leq u(t) \leq 1$  where

$$-\frac{\pi}{2} < \arctan x < \frac{\pi}{2}$$

Here

$$g(x) = \arctan x, g(0) = 0, g'(x) = \frac{1}{1+x^2} \leq 1$$

*Remark*—In the next section an example is considered in which it is shown that the switching boundary is a single-valued curve and the switching boundary is constructed. It appears that the special hypothesis is easily verified experimentally.

## Example and Conclusion

Consider the Duffing equation with control

$$\ddot{x} + x + 2x^3 = u(t)$$

or the system

$$\dot{x} = y$$

$(S_d)$

$$\dot{y} = -x - 2x^3 + u(t)$$

with  $-1 \leq u(t) \leq 1$ .

For the time optimal problem  $u(t) = \pm 1$  on various intervals of time, thus consider

$$(S_{\pm}) \quad \begin{aligned} \dot{x} &= y \\ \dot{y} &= -x - 2x^3 \pm 1 \end{aligned}$$

Also consider the adjoint system

$$(A) \quad \dot{\eta}_1 = -\frac{\partial H}{\partial x} = \eta_2(1 + 6x^2(t))$$

$$\dot{\eta}_2 = -\frac{\partial H}{\partial y} = -\eta_1$$

from which is found  $\eta_2(t)$ , that determines the interval of time that the solutions of  $S_{+}$  and  $S_{-}$  are followed to obtain the extremal responses.



The switching boundary,  $W$ , is now constructed as in the previous theory.  $W_+^1$  is the solution of  $S_+$  through the origin and lying in the fourth quadrant, see Figure 1.  $W_-^1$  is the solution of  $S_-$  through the origin and lying in the second quadrant.

$W_-^1$  is now reflected along solutions of  $S_-$ . From the points 1, 2, ..., 9 of  $W_-^1$  follow backwards in time along the corre-

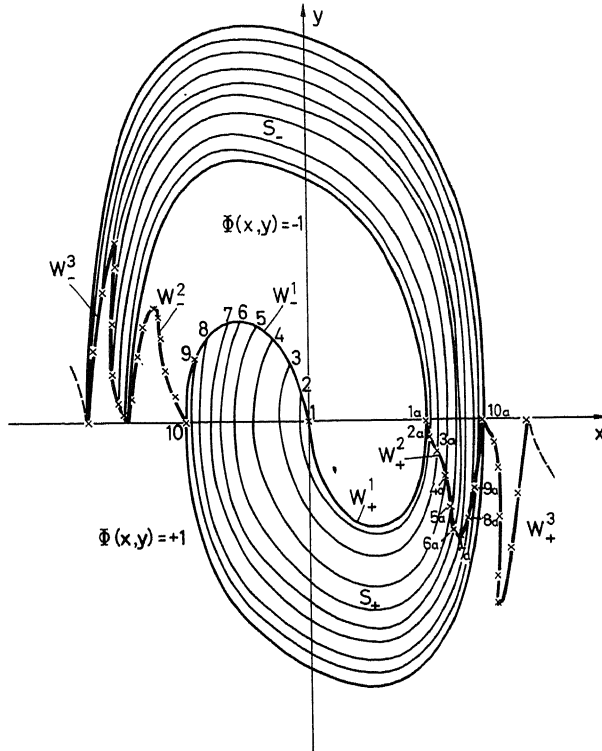


Figure 1

sponding solutions of  $S_-$  for a time duration equal to the interval between zeros of  $\eta_2(t)$ , i.e. from, say, point 1 with  $\eta_1(0) = 1$ ,  $\eta_2(0) = 0$  follow  $S_-$  and consider the response of the adjoint system (A) until  $\eta_2(t)$  is again zero. This determines a point 1a on  $W_+^2$ . This is repeated for each of the points 2, 3, ..., 9 and gives the corresponding points 2a, 3a, ..., 9a of  $W_+^2$ . Since

it is known that  $W_+^2$  is a  $C^1$  curve, the values between 2a, 3a, ..., 8a can be filled in by extrapolation. For this reason only a finite number of extremal responses need be considered in generating  $W$ .

Now that  $W_+^2$  has been found,  $W_-^2$  is the reflection through the origin (see Figure 1).  $W_+^2$  is formed by reflecting  $W_+^1$  along a conjugate interval as defined above. The remaining segments of  $W$  are built up in exactly the same manner.

Thus the feedback controller  $u = \Phi(x, y)$  for the Duffing equation ( $S_d$ ) is found. It is very easy to store this function and therefore the synthesis is complete.  $\Phi(x, y) = +1$  below the switching boundary  $W$  and  $\Phi(x, y) = -1$  above the switching boundary  $W$ .

There is some question as to the accuracy with which  $W$  has been determined in Figure 1, but if it is nearly correct, then it is seen that  $W = W(x)$  is indeed a single-valued function over the two flows  $S_+$  and  $S_-$ .

## References

- 1 BOLTYANSKII, V.G., GAMKRELIDZE, R.V., and PONTRYAGIN, L.S. The theory of optimal processes I, the Maximum Principle. *Bull. Acad. Sci. U.R.S.S.* 24 (1960) 3
- 2 BUSHAW, D. W. Optimal discontinuous forcing terms. *Contrib. to Nonlinear Oscillations IV*, 29. 1958. Princeton
- 3 HARTMAN, P. A lemma in the theory of structural stability of differential equations. *Proc. Amer. math. Soc.* 11 (1960) 610
- 4 HARVEY, C. A. and LEE, E. B. On the uniqueness of time optimal control for linear processes. *J. Math. Anal. & App.* 5 (1962)
- 5 LASALLE, J. P. The time optimal control problem. *Nonlinear Oscillations V*. 1961. Princeton
- 6 LEE, E. B. and MARKUS, L. Synthesis of optimal control for Nonlinear processes with one degree of freedom. *Symp. Nonlinear Vibrations*. Sept. 1961. Kiev
- 7 LEE, E. B. and MARKUS, L. On the existence of optimal controls. *Trans. Amer. Soc. mech. Engrs. J. Basic Engng* (1962) 13
- 8 LEE, E. B., and MARKUS, L. Optimal control for nonlinear processes. *Arch. Rational Mech. & Anal.* 8 (1961) 36
- 9 MARKUS, L. Global structure of ordinary differential equations in the plane. *Trans. Amer. math. Soc.* 76 (1964) 127
- 10 MARKUS, L., and YAMABE, H. Global stability criteria for differential systems. *Osaka math. J.* 12 (1960) 305
- 11 PONTRYAGIN, L. S. Optimal control processes. *Progr. meth. Sci., Moscow* 14 (1959) 3

## DISCUSSION

Y. SAKAWA, *Department of Electrical Engineering, Kyoto University, Kyoto, Japan*

We obtained numerically the optimum switching curve of a non-linear system governed by

$$\ddot{x} + \varepsilon(x^2 - 1)\dot{x} + x = u, \quad \varepsilon > 0, \quad |u| \leq \frac{1}{2}$$

by using an analogue computer, as mentioned in the paper by Hayashi and myself, and we recognized that when the value of  $\varepsilon$  is small the domain of controllability covers the whole phase plane, but when the

value of  $\varepsilon$  is great the domain of controllability is bounded in the phase plane. In this case the condition

$$\frac{\partial f(x, y)}{\partial y} \geq 0 \quad [f(x, y) = \varepsilon(x^2 - 1)y + x] \text{ in } R^2$$

is not realized

Did the authors obtain any general analytical results for such a system?

E. B. LEE and L. MARKUS, *in reply*

See Theorem 7 of Reference 8 of our paper for a discussion of this case.

# On Optimal Control of Systems with Multi-norm Constraints\*

P. E. SARACHIK and G. M. KRANC

## Summary

In this paper the optimal time problem is considered for multiple input systems when each input may be subject to a separate and possibly different type of constraint.

Specifically the control of a plant described by an  $n$ th order, linear, time varying differential equation and which has  $r$  inputs and  $m$  outputs is considered. Each input  $u_i(t)$  is constrained by a condition of the form

$$\left[ \int_0^T |u_i(t)|^{p_i} dt \right]^{\frac{1}{p_i}} \leq L_i, \text{ for } i = 1, 2, \dots, r$$

where  $p_i \geq 1$ . Note that the constraints are applied to each input separately so that this problem includes those practical cases when some inputs may be limited in magnitude, some in energy, and others in still different ways. The problem is to find the set of inputs  $u_i(t)$  which satisfy the constraints and which will make the plant outputs equal to a set of desired outputs in minimum time  $T$ .

The technique presented in this paper for solving the problem is an extension of the methods of functional analysis first used by Krasovskii and Kulikowski to solve optimal time problems with a single constrained input. A useful feature of the solution presented in this paper is that a very general norm is defined for the input vector  $u(t) = [u_1(t), u_2(t), \dots, u_r(t)]$  which incorporates all the separate constraints into a single constraint condition. This permits a solution of the optimal time problem for this case, which may be obtained with no more formal difficulty than in the single input case.

## Sommaire

Cette communication a trait au problème du temps optimal pour des systèmes à entrées multiples dans lesquels chaque entrée peut être séparément l'objet d'un type donné de contrainte, éventuellement différent pour chacune d'elles.

Elle s'attache en particulier au réglage d'une installation qui est régie par une équation différentielle linéaire d'ordre  $n$  par rapport au temps et qui possède  $r$  entrées et  $m$  sorties. Chaque entrée  $u_i(t)$  est soumise à une contrainte qui s'exprime par une condition de la forme

$$\left[ \int_0^T |u_i(t)|^{p_i} dt \right]^{\frac{1}{p_i}} \leq L_i, \text{ pour } i = 1, 2, \dots, r$$

où  $p_i \geq 1$ . Il convient de remarquer que ces contraintes sont appliquées séparément à chaque entrée, de sorte que le problème ainsi posé inclut les cas pratiques où certaines entrées peuvent être limitées en amplitude, certaines en énergie et d'autres de manière encore différente. Le problème consiste à trouver l'ensemble des entrées  $u_i(t)$  qui, tout en satisfaisant les conditions représentant les contraintes, rende les sorties égales à l'ensemble des sorties désirées, et cela en un temps minimal  $T$ .

La technique développée dans cette communication pour résoudre ce problème est une extension des méthodes d'analyse fonctionnelle appliquées pour la première fois par Krasovskii et Kulikowski à la résolution des problèmes de temps optimal dans lesquels il n'y a qu'une seule entrée sujette à contrainte. Un des avantages offerts par la méthode présentée dans cette communication réside dans le fait qu'elle permet de définir une norme très générale pour le vecteur d'entrée  $u(t) = [u_1(t), u_2(t), \dots, u_r(t)]$  ce qui permet d'englober en

une unique condition de contrainte toutes les contraintes individuelles. D'où la possibilité d'obtenir pour ce cas une solution du problème du temps optimal, et cela sans plus de difficulté formelle que pour le cas de l'entrée unique.

## Zusammenfassung

Der Aufsatz betrachtet das Problem der zeitoptimalen Systeme mit mehreren Eingängen, wobei jeder Eingang einer getrennten und möglicherweise unterschiedlichen Art von Beschränkungen (Bedingungen) unterliegen kann.

Im besonderen wird die Regelung einer Strecke mit  $r$  Eingängen und  $m$  Ausgängen betrachtet, die sich durch eine lineare Differentialgleichung  $n$ -ter Ordnung mit zeitvariablen Koeffizienten beschreiben läßt. Jeder Eingang  $u_i(t)$  unterliegt einer Bedingung der folgenden Form:

$$\left[ \int_0^T |u_i(t)|^{p_i} dt \right]^{\frac{1}{p_i}} \leq L_i, \text{ für } i = 1, 2, \dots, r$$

wobei  $p_i \geq 1$  ist. Da diese Bedingungen für jeden Eingang einzeln gelten, schließt das hier behandelte Problem auch die praktischen Fälle mit ein, bei denen in einigen Eingängen der Betrag, in einigen die Energie, und andere in ganz anderer Weise begrenzt sein können. Es geht darum, die Gruppe von Eingängen  $u_i(t)$  zu finden, die den Beschränkungen genügt, und die in einer minimalen Zeit  $T$  die Regelgröße auf die gewünschten Werte bringt.

Die hier zur Lösung des Problems benutzte Methode ist eine Erweiterung der Methoden der Funktionalanalysis, wie sie zuerst Krasovskii und Kulikowski für zeitoptimale Systeme verwendeten, bei denen nur ein Eingang einer Beschränkung unterliegt. Eine nützliche Eigenschaft der hier vorgelegten Lösung liegt in der Definition einer sehr allgemeinen Norm für den Eingangsvektor  $u(t) = [u_1(t), u_2(t), \dots, u_r(t)]$ , die alle getrennten Bedingungen in einer einzigen erfaßt. Hierdurch läßt sich das vorliegende zeitoptimale Problem ohne zusätzliche formelle Schwierigkeiten wie im Falle eines Systems mit einem Eingang lösen.

## Introduction

In a recent paper<sup>1</sup>, the authors showed how Kulikowski's<sup>2,3</sup> application of functional analysis techniques to the solution of optimal control problems could be extended to multiple input systems. This was accomplished by defining a general norm on the entire input vector so that constraint conditions could be applied to all the inputs simultaneously. That paper, however, was concerned with problems in which all inputs have the same type of constraint (e.g., when all inputs are limited in magnitude or all are limited in power or energy), and the constraint is imposed on the totality of the inputs, (e.g., when the combined power dissipated by all inputs is limited).

In the present paper, a similar procedure is applied to a more general problem. In this case a problem is considered, in which each input is individually constrained and where the individual constraints need not be of the same type. This problem is approached by first embedding it in a still more general problem

\* This research was supported under AF-AFOSR-62-144 and NSF G-14514.

in which a very general norm constraint is applied on the entire input vector. It will then be easy to show that the solution to the actual problem as well as to the problems considered elsewhere<sup>1,2</sup> are just special cases of this general solution.

### Statement of the Problem

This paper is concerned with the solution of time optimal problems when separate constraints are imposed on each input. Given:

(a) A completely output controllable plant (see Appendix I) described by the linear differential equations

$$\dot{x}(t) = F(t)x(t) + D(t)u(t) \quad (1)$$

with output  $y(t) = M(t)x(t)$ , where the state of the system at any time  $t$  is represented by the  $n$ -dimensional vector  $x(t)$ ; the  $r$  control inputs are represented by the  $r$ -dimensional vector  $u(t)$ ; the  $n \times n$  matrix  $F(t)$ , the  $n \times r$  matrix  $D(t)$  and the  $m \times n$  matrix  $M(t)$  are called the system matrix, distribution matrix and output matrix respectively;

(b) The initial state of the plant at time  $t_0$

$$x(t_0) = x_0$$

(c) A desired output  $y^d(t)$ ;

(d) A separate constraint on each plant input  $u_i(t)$  in the form

$$\|u_i\|_{p_i} \triangleq \left[ \int_{t_0}^{t_1} |u_i(\tau)|^{p_i} d\tau \right]^{1/p_i} \leq L_i \quad (2)$$

where  $p_i \geq 1$  for  $i = 1, 2, \dots, r$ .

(Note that in this problem the  $p_i$  and  $L_i$  may be different for each  $i$ . This means that each component of the control vector (each control input) may have a different type of constraint. For example, if the input  $u_j$  is to be constrained in magnitude then  $p_j = \infty$ ; if it is to be constrained in energy then  $p_j = 2$ , etc.)

The problem is to find that control  $u_0(t)$  which satisfies the constraints and makes the plant output  $y$  at time  $t_1$  equal to the desired output (i. e.,  $y(t_1) = y^d(t_1)$ ) for minimum elapsed time  $T$  defined by  $T = t_1 - t_0 \geq 0$ .

### A Simplified Problem

It is instructive to first consider a simplified version of the general problem. Assume a time invariant, linear plant which can be activated by two possible inputs  $u_1(t)$  and  $u_2(t)$ , and which has a single output. If  $x(t)$  is the response (the total output consists of the response  $x(t)$  and the response due to initial conditions which is assumed to be known) of this plant to the signals  $u_1(t)$  and  $u_2(t)$  which are applied at  $t = 0$  then

$$x(t) = \int_0^t k_1(t-\tau)u_1(\tau)d\tau + \int_0^t k_2(t-\tau)u_2(\tau)d\tau \quad (3)$$

where  $k_1(t)$  and  $k_2(t)$  are appropriate impulse responses.

The problem is to determine  $u_1(t)$  and  $u_2(t)$  so that the response  $x(t)$  should reach a desired value  $X$  in minimum time  $t = T_0$  subject to the constraint  $\|u_1\|_{p_1} \leq L_1$  and  $\|u_2\|_{p_2} \leq L_2$  [see eqn (2)].

First, let  $x(t) = X$  at some given time  $T$ . Then

$$X = x(T) = x_1(T) + x_2(T) \quad (4)$$

where

$$x_1(T) = \int_0^T k_1(T-\tau)u_1(\tau)d\tau \quad (5)$$

and

$$x_2(T) = \int_0^T k_2(T-\tau)u_2(\tau)d\tau \quad (6)$$

It follows from Hölder's inequality [see Appendix II, eqn (84)] applied to eqns (5) and (6) that

$$|x_1(T)| = \left| \int_0^T k_1(T-\tau)u_1(\tau)d\tau \right| \leq \|k_1\|_{q_1} \|u_1\|_{p_1} \quad (7)$$

and

$$|x_2(T)| = \left| \int_0^T k_2(T-\tau)u_2(\tau)d\tau \right| \leq \|k_2\|_{q_2} \|u_2\|_{p_2} \quad (8)$$

where

$$\|k_i\|_{q_i} \triangleq \left[ \int_0^T |k_i(t)|^{q_i} dt \right]^{1/q_i} \quad (9)$$

and  $1/p_i + 1/q_i = 1$ .

Simplify the notation denoting  $\|u_i\|_{p_i}$  and  $\|k_i\|_{q_i}$  by  $\|u_i\|$  and  $\|k_i\|$  respectively. By making use of the constraints on  $u_i(t)$  and eqns (7) and (8)

$$\frac{|x_1(T)|}{\|k_1\|} \leq \|u_1\| \leq L_1 \quad (10)$$

and

$$\frac{|X - x_1(T)|}{\|k_2\|} = \frac{|x_2(T)|}{\|k_2\|} \leq \|u_2\| \leq L_2 \quad (11)$$

or rearranging the above equations,

$$L_1 \|k_1\| \geq |x_1(T)| \quad (12)$$

and

$$L_2 \|k_2\| \geq |X - x_1(T)| \quad (13)$$

The smallest value of  $T$  for which both inequalities (12) and (13) are satisfied is the required optimal time  $T_0$ .

Assume now that for some fixed value of  $x_1(T)$ , say  $x_1$ , there is a range of values of  $T$  for which both eqns (12) and (13) are satisfied. When  $x_1 = 0$ , this occurs when  $T \geq T_1$  (see Figure 1); when  $x_1 \neq 0$  and  $\text{sgn } x_1 = \text{sgn } X$ , inequalities (12) and (13) are both satisfied when  $T_1 > T \geq T_0$ .  $T_0$  is then the optimal time and it can be found by requiring that both inequalities (12) and (13) are satisfied simultaneously with a sign of equality when  $\text{sgn } x_1 = \text{sgn } X$ . (Note: when  $x_1 = x_1' \neq 0$  and  $\text{sgn } x_1' \neq \text{sgn } X$ , then inequalities (12) and (13) may be also satisfied as shown in Figure 1 for some  $T \geq T_2$  but the minimum value of  $T_2$  is always greater than  $T_1$  and therefore also greater than  $T_0$ .)

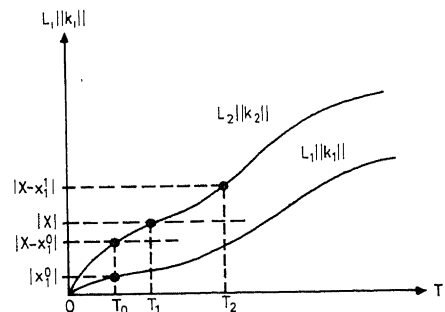


Figure 1. Possible plots of  $L_i \|k_i\|$  versus  $T$

Therefore,  $L_1 \|k_1\| = |x_1^0|$  (14)

and  $L_2 \|k_2\| = |X - x_1^0|$  (15)

and also  $x_1(T_0) = x_1^0 = L_1 \|k_1\| \operatorname{sgn} X$  (16)

Using a similar argument as above it is found that

$$x_2(T_0) = X - x_1^0 = L_2 \|k_2\| \operatorname{sgn} X \quad (17)$$

Assume now that eqns (14) and (15) are satisfied. This implies that inequalities (10) and (11) and therefore also eqns (7) and (8) must be satisfied with a sign of equality. By Hölder's inequality [see Appendix II, eqns (84) and (85)] applied to eqns (5) and (6) this occurs if

$$u_1(\tau) = \frac{x_2(T)}{(\|k_1\|_{q_1})^{q_1}} |k_1(T-\tau)|^{q_1-1} \operatorname{sgn} k_1(T-\tau) \quad (18)$$

and

$$u_2(\tau) = \frac{x_2(T)}{(\|k_2\|_{q_2})^{q_2}} |k_2(T-\tau)|^{q_2-1} \operatorname{sgn} k_2(T-\tau) \quad (19)$$

Optimal inputs  $u_1^0$  and  $u_2^0$  are obtained by substituting  $T = T_0$  in eqns (18) and (19) and by making use of eqns (16) and (17). Thus

$$u_1^0(\tau) = L_1 (\|k_1\|_{q_1})^{1-q_1} |k_1(T_0-\tau)|^{q_1-1} \operatorname{sgn} \frac{k_1(T_0-\tau)}{X} \quad (20)$$

and

$$u_2^0(\tau) = L_2 (\|k_2\|_{q_2})^{1-q_2} |k_2(T_0-\tau)|^{q_2-1} \operatorname{sgn} \frac{k_2(T_0-\tau)}{X} \quad (21)$$

To simplify the calculation of  $T_0$ , note, making use of eqns (14) and (15), that

$$L_1 \|k_1\| + L_2 \|k_2\| = |x_1^0| + |X - x_1^0| \geq |X| \quad (22)$$

Clearly then,  $T_0$  is the least value of  $T$  which satisfies the equation

$$L_1 \|k_1\|_{q_1} + L_2 \|k_2\|_{q_2} = |X| \quad (23)$$

In fact, if eqn (23) cannot be satisfied for any  $T$ , then there is no solution to the problem.

Making use of eqn (23), eqns (20) and (21) can be written

$$u_1^0(\tau) = \frac{XL_1 (\|k_1\|_{q_1})^{1-q_1}}{L_1 \|k_1\|_{q_1} + L_2 \|k_2\|_{q_2}} |k_1(T_0-\tau)|^{q_1-1} \operatorname{sgn} k_1(T_0-\tau) \quad (24)$$

$$u_2^0(\tau) = \frac{XL_2 (\|k_2\|_{q_2})^{1-q_2}}{L_1 \|k_1\|_{q_1} + L_2 \|k_2\|_{q_2}} |k_2(T_0-\tau)|^{q_2-1} \operatorname{sgn} k_2(T_0-\tau) \quad (25)$$

It can be verified by direct substitution that the inputs given in eqns (24) and (25) satisfy eqn (3) for any time  $t = T$  (not necessarily for  $T = T_0$ ) and that  $x(T) = X$ , the desired value. At the same time it can also be verified that

$$\|u_1\|_{p_1} = \frac{|x_1|}{\|k_1\|_{q_1}} \quad (26)$$

$$\|u_2\|_{p_2} = \frac{|X - x_1|}{\|k_2\|_{q_2}} = \frac{|x_2|}{\|k_2\|_{q_2}} \quad (27)$$

$$\text{and} \quad \frac{\|u_1\|_{p_1}}{\|u_2\|_{p_2}} = \frac{L_1}{L_2} \quad (28)$$

It follows from inequalities (10) and (11) that eqns (26) and (27) give minimum values for  $\|u_i\|_{p_i}$ . Thus it is seen that the time optimal problem could have been approached by first trying to find  $u_1(\tau)$  and  $u_2(\tau)$  which satisfy the terminal condition  $x(T) = X$  with minimum values  $\|u_i\|_{p_i}$  and also the additional constraint of eqn (28).

### A General Norm<sup>5</sup>

Before proceeding to the general solution of the problem it will be convenient at this point to consider a means of representing the set of  $r$  constraints (2) as a single constraint condition. This is readily done by defining a norm of  $u$  as

$$\|u\| = \max_i [\|u_i\|_{p_i}/L_i] \quad (29)$$

Note that when the single condition

$$\|u\| \leq 1 \quad (30)$$

is satisfied, then it is apparent that the  $r$  conditions given by eqn (2) are also satisfied. This means that the  $r$  input constraints can be imposed by requiring that the single condition eqn (30) be satisfied. Unfortunately, the norm defined in eqn (29) is a difficult one to work with directly; a simplification in the solution is achieved by a further embedding of the constraint conditions in a still more general form of norm on  $u(t)$ .

This more general norm can be defined by

$$\|u\|_p \triangleq \left[ \sum_{i=1}^r L_i^{-p} \|u_i\|_{p_i}^p \right]^{1/p} \quad (31)$$

where  $p \geq 1$  and  $\|u_i\|_{p_i}$  is defined in eqn (2). Using the same method as Kirillova<sup>6</sup> it can be shown that in the limit as  $p \rightarrow \infty$  the solution with  $\|u\|_p$  constrained approaches the solution with  $\|u\|$  constrained, so that the constraints of eqn (2) will be replaced by the constraint

$$\|u\|_p \leq 1 \quad (32)$$

The solution of the originally stated problem is thus obtained by first solving the more general problem which has a constraint given by eqn (32) and then letting  $p \rightarrow \infty$ .

It may be of interest to point out that the problem considered by Kranc and Sarachik<sup>1</sup> is a special case of this more general problem. This fact is easily verified by setting all  $p_i = p$  and all  $L_i = L$  so that condition (32) gives

$$\|u\|_p = \frac{1}{L} \left[ \int_{t_0}^{t_1} \sum_{i=1}^r |u_i(\tau)|^p d\tau \right]^{1/p} \leq 1 \quad (33)$$

The problem considered<sup>1</sup> was the same as that stated here except that the constraints were those of eqn (33) instead of eqn (2).

### Solution of the Problem

In order to solve the stated problem, utilizing the approach of Krasovskii<sup>7</sup>, consider first the following alternate problem:

Given a plant, initial condition, and desired output as described in Problem Statement (a), (b) and (c).

Find that input  $u(t)$  which at a fixed time  $t_1$  makes  $y(t_1) = y^d(t_1)$  and which has a minimum norm  $\|u\|_p$ .

Note that in this alternate problem the time at which the actual plant output must equal the desired output is fixed at the start and no constraints are imposed. The minimum time solution will be obtained from this 'minimum norm' solution by choosing the smallest time for which the minimum norm  $\|u\|_p$  is just equal to unity.

The solution of the differential equations describing the plant (1) which satisfy the given initial conditions  $x_0$  at time  $t_0$  is

$$x(t) = \Phi(t, t_0) x_0 + \int_{t_0}^t \Phi(t, \tau) D(\tau) u(\tau) d\tau \quad (34)$$

where  $\Phi(t, \tau)$  is a fundamental matrix<sup>8</sup> of the plant which also satisfies  $\Phi(t, t) = I$  and will be called the transition matrix.

The output is given by

$$y(t) = M(t) \Phi(t, t_0) x_0 + \int_{t_0}^t M(t) \Phi(t, \tau) D(\tau) u(\tau) d\tau \quad (35)$$

These equations become simpler if the following definitions are introduced

$$e(t) \triangleq y(t) - M(t) \Phi(t, t_0) x_0 \quad (36)$$

and

$$H(t, \tau) \triangleq M(t) \Phi(t, \tau) D(\tau) \quad (37)$$

Note that the  $m \times r$  matrix  $H(t, \tau)$  is the impulse response matrix of the system and  $e(t)$  represents the difference between the actual output and the output which would have been obtained from the initial condition alone if no inputs had been applied for  $t > t_0$ .

Using eqns (36) and (37), eqn (35) becomes

$$e(t) = \int_{t_0}^t H(t, \tau) u(\tau) d\tau \quad (38)$$

The problem now is to find the input vector  $u(\tau)$  with  $\min \|u\|_p$  which satisfies

$$\int_{t_0}^{t_1} H(t_1, \tau) u(\tau) d\tau = e^d(t_1) \quad (39)$$

where

$$e^d(t_1) = y^d(t_1) - M(t_1) \Phi(t_1, t_0) x_0$$

or equivalently to find the input vector which satisfies the  $m$  equations

$$\int_{t_0}^{t_1} h_j(t_1, \tau) u(\tau) d\tau = e_j^d(t_1) \quad \text{for } j = 1, 2, \dots, m \quad (40)$$

where  $h_j(t, \tau)$  is the  $j$ th row of the matrix  $H(t, \tau)$  and  $e_j^d(t_1)$  is the  $j$ th element of the vector  $e^d(t_1)$ .

Since the plant is completely output-controllable it is known that at least one input exists which meets these terminal conditions for some  $t_1$  and the one with minimum  $\|u\|_p$  is sought.

Now denote by  $U_A$  the set of all inputs which satisfy eqn (40). Define the functional  $f_A(h_j)$  by

$$f_A(h_j) \triangleq \int_{t_0}^{t_1} h_j(t_1, \tau) u(\tau) d\tau \quad (41)$$

where  $u(\tau)$  belongs to  $U_A$  ( $u(\tau) \in U_A$ ). Therefore

$$f_A(h_j) = e_j^d(t_1) \quad (42)$$

and since  $f_A(\cdot)$  is a linear functional, if any linear combination of the  $h_j(t, \tau)$  is considered, such as

$$k(t_1, \tau) = \sum_{j=1}^m \lambda_j h_j(t_1, \tau) = \lambda H(t_1, \tau) \quad (43)$$

where  $\lambda$  is an  $m$ -dimensional row vector, then by linearity and eqns (42) and (43)

$$f_A(k) = f_A\left(\sum_{j=1}^m \lambda_j h_j\right) = \sum_{j=1}^m \lambda_j f_A(h_j) = \lambda e^d(t_1) \quad (44)$$

for any  $k(t_1, \tau)$  which can be represented by eqn (43).

Defining the quantities

$$\|k_i\|_{q_i} \triangleq \left[ \int_{t_0}^{t_1} |k_i(t_1, \tau)|^{q_i} d\tau \right]^{1/q_i} \quad (45)$$

and

$$\|k\|_q \triangleq \left[ \sum_{i=1}^r L_i^q \|k_i\|_{q_i}^q \right]^{1/q} \quad (46)$$

where  $k_i(t_1, \tau)$  is the  $i$ th component of the row vector  $k(t_1, \tau)$  defined in eqn (43) and  $q_i$  and  $q$  are related to  $p_i$  and  $p$  of eqn (31) by  $1/q_i + 1/p_i = 1$  and  $1/p + 1/q = 1$ .

Now consider the quantity  $\|f_A\|$  (called the norm of the functional  $f_A$ ) defined by

$$\|f_A\| \triangleq \max_k \left\{ \frac{f_A(k)}{\|k\|_q} \right\} \quad (47)$$

where  $k$  is given by eqn (43). Using eqns (43) and (44), eqn (47) becomes

$$\|f_A\| = \max_{\lambda} \left\{ \frac{|\lambda e^d(t_1)|}{\left\| \sum_{j=1}^m \lambda_j h_j \right\|_q} \right\} = \max_{\lambda e^d=1} \left\{ \frac{1}{\left\| \sum_{j=1}^m \lambda_j h_j \right\|_q} \right\} \quad (48)$$

Denote by  $\lambda^*$  the vector whose components  $\lambda_1^* \dots \lambda_m^*$  minimize  $\left\| \sum_{j=1}^m \lambda_j h_j \right\|_q$  subject to the condition  $\lambda e^d(t_1) = 1$ . Therefore

$$\min_{\lambda e^d=1} \left\| \sum_{j=1}^m \lambda_j h_j \right\|_q = \left\| \sum_{j=1}^m \lambda_j^* h_j \right\|_q \triangleq \|k^*\|_q \quad (49)$$

and eqn (48) can be written as

$$\|f_A\| = \frac{1}{\|k^*\|_q} \quad (50)$$

Considering now the functional

$$f_A(k^*) = \int_{t_0}^{t_1} k^*(t_1, \tau) u(\tau) d\tau \quad (51)$$

and applying the generalization of Hölder's inequality derived in Appendix II it is found that

$$|f_A(k^*)| \leq \|k^*\|_q \|u\|_p \quad (52)$$

or by eqn (50) and because  $f_A(k^*) = \lambda^* e^d(t_1) = 1$ , that

$$\|u\|_p \geq \frac{|f_A(k^*)|}{\|k^*\|_q} = \frac{1}{\|k^*\|_q} = \|f_A\| \quad (53)$$

Inequality (53) states the necessary condition that  $u(\tau) \in U_A$ .

In addition, from eqn (53) is obtained the necessary condition that  $\mathbf{u}(\tau) \in U_A$  with a min norm  $\|\mathbf{u}\|_p$ , namely

$$\|\mathbf{u}\|_p = \frac{1}{\|\mathbf{k}^*\|_q} = \|f_A\| \quad (54)$$

The above condition will be satisfied whenever inequality (53) and therefore inequality (52) are satisfied with the sign of equality. By Appendix II, when the vector  $\mathbf{u}(\tau)$  has its elements  $u_i(\tau)$  of the form

$$u_i(\tau) = K L_i^q \|k_i^*\|_{q_i}^{q-q_i} |k_i^*(t_1, \tau)|^{q_i-1} \text{sgn } k_i^*(t_1, \tau) \quad (55)$$

then the sign of equality holds for eqn (52).

Substituting eqn (55) into (54) the constant  $K$  is evaluated to be

$$K = \frac{1}{(\|\mathbf{k}^*\|_q)^q} \quad (56)$$

Thus, eqn (54) is satisfied if and only if

$$u_i(\tau) = \frac{L_i^q}{(\|\mathbf{k}^*\|_q)^q} \|k_i^*\|_{q_i}^{q-q_i} |k_i^*(t_1, \tau)|^{q_i-1} \text{sgn } k_i^*(t_1, \tau) \quad (57)$$

It follows from the above argument that the necessary condition that  $\mathbf{u} \in U_A$  with a minimum norm  $\|\mathbf{u}\|_p$  is that  $u_i(\tau)$  be of the form dictated by eqn (57). To show the sufficiency of this proposition it must be demonstrated that the signal  $\mathbf{u}(\tau)$  given by eqn (57) is member of  $U_A$  in that it satisfies eqn (40). This is done in Appendix III using an indirect argument.

As was mentioned earlier, the constraints of eqn (2) on the input components in the original problem can be imposed merely by considering the norm  $\|\mathbf{u}\|$  as defined in eqn (29). In accordance with this earlier discussion the input is found which has minimum norm  $\|\mathbf{u}^*\|$  and which satisfies the condition  $\mathbf{y}(t_1) = \mathbf{y}^d(t_1)$  by just setting  $q = 1$  (this corresponds to  $p = \infty$ ) in eqn (57). In this case it is found that

$$u_i(t)^* = \frac{L_i}{\|\mathbf{k}^*\|_1} \|k_i^*\|_{q_i}^{1-q_i} |k_i^*(t_1, t)|^{q_i-1} \text{sgn } k_i^*(t_1, t) \quad (58)$$

where the starred quantities are obtained by finding  $\lambda^*$  which gives

$$\|\mathbf{k}^*\|_1 = \min_{\lambda} \sum_{i=1}^r L_i \|k_i\|_{q_i} \quad (59)$$

subject to the constraint that  $\lambda e^d(t_1) = 1$ .

It should be noted that the actual minimization required by eqn (59) is much more difficult than for single norm constraints.

To obtain the time optimal solution required by the original statement of the problem the effect of the constraints must be considered. It has been shown that all the constraints are satisfied by the single condition

$$\|\mathbf{u}\| \leq 1 \quad (60)$$

Now from eqn (52) it is seen that for  $p = \infty$ ,  $q = 1$

$$\|\mathbf{u}^*\| \geq \frac{1}{\|\mathbf{k}^*\|_1} \quad (61)$$

This shows that for any  $t_1$  a solution is possible if and only if

$$\|\mathbf{k}^*\|_1 \geq 1 \quad (62)$$

When  $\|\mathbf{k}^*\|_1$  is a continuous function of  $t_1$  the minimum elapsed time  $T_0 = t_1^* - t_0$  is achieved for the smallest  $t_1 = t_1^*$  which makes

$$\|\mathbf{k}^*\|_1 = 1 \quad (63)$$

This equation is used to find  $t_1^*$ . Now if the minimum norm solution of eqn (58) is used then for  $t_1 = t_1^*$  it is seen that

$$\|\mathbf{u}^*\| = \frac{1}{\|\mathbf{k}^*\|_1} = 1 \quad (64)$$

so the solution given by eqn (58) satisfies the constraints and the end conditions for minimum elapsed time and is therefore the time-optimal solution. Using eqn (64) this gives

$$u_i(t)_{\text{opt}} = L_i \|k_i^*\|_{q_i}^{1-q_i} |k_i^*(t_1^*, t)|^{q_i-1} \text{sgn } k_i^*(t_1^*, t) \quad (65)$$

where  $k_i^*$  is obtained by solving eqn (59) for  $\lambda^*$  and substituting this value into eqn (43). Eqn (63) must then be solved for minimum  $t_1$  which gives  $t_1^*$ .

It can be verified by forming  $\|u_i(t)_{\text{opt}}\|_{p_i}$  from eqn (2) for the input given by eqn (65) that  $\|u_i(t)_{\text{opt}}\|_{p_i} = L_i$  for all  $i$ . This shows that the optimal time solution is given by an input for which all constraints are satisfied with an equality.

If  $\|\mathbf{k}^*\|_1$  is not a continuous function of  $t_1$  the necessary and sufficient condition for the existence of a solution is still eqn (62). In this case, however, a true minimum may not exist, although the greatest lower bound (infimum) may be approached arbitrarily closely.

### Numerical Example

To illustrate how the preceding results may be applied, a very simple example will be considered taking the plant shown in Figure 2. Its differential equations written in normal form are

$$\begin{bmatrix} \dot{x}_1(t) \\ \dot{x}_2(t) \\ \dot{x}_3(t) \end{bmatrix} = \begin{bmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \\ 0 & 0 & 0 \end{bmatrix} \begin{bmatrix} x_1(t) \\ x_2(t) \\ x_3(t) \end{bmatrix} + \begin{bmatrix} 0 & 0 \\ 0 & 1 \\ 1 & 0 \end{bmatrix} \begin{bmatrix} u_1(t) \\ u_2(t) \end{bmatrix} \quad (66)$$

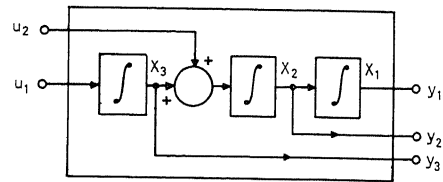


Figure 2. Plant for numerical example

This system is time invariant and its fundamental matrix  $\Phi(t - t_0)$  is obtained from the solution of the unforced equations which satisfy the initial condition  $\mathbf{x}(t_0) = \mathbf{x}_0$ . For invariant systems the easiest method of obtaining the fundamental matrix  $\phi(t - t_0)$  is by using Laplace transforms. For a system described by

$$\dot{\mathbf{x}} = \mathbf{F}\mathbf{x} + \mathbf{D}\mathbf{u}$$

$\Phi(t)$  is found from

$$\phi(t) = \mathcal{L}^{-1} \{ [s\mathbf{I} - \mathbf{F}]^{-1} \}$$

where  $\mathcal{L}^{-1}$  means inverse Laplace transformation.

Applying this procedure to the system in eqn (66) gives

$$[sI - F]^{-1} = \begin{bmatrix} s & -1 & 0 \\ 0 & s & -1 \\ 0 & 0 & s \end{bmatrix}^{-1} = \frac{1}{s^3} \begin{bmatrix} s^2 & s & 1 \\ 0 & s^2 & s \\ 0 & 0 & s^2 \end{bmatrix}$$

Taking the inverse  $\mathcal{L}$  transform of each element yields

$$\Phi(t) = \begin{bmatrix} 1 & t & \frac{t^2}{2} \\ 0 & 1 & t \\ 0 & 0 & 1 \end{bmatrix} \quad (67)$$

To make the example specific say that it is desired to find the inputs  $u_1(t)$  and  $u_2(t)$  which will take the system from a given initial state to the origin  $x^d = 0$  in minimum time. The input  $u_1(t)$  is constrained in energy ( $\|u_1\|_2 \leq L_1$ ) and the input  $u_2(t)$  is constrained in amplitude ( $\|u_2\|_\infty \leq L_2$ ). Note that in this case since the output  $y(t) = x(t)$  which is the state vector itself,  $M$  in eqn (1) is the identity matrix.

From eqn (37)  $H(t_1 - \tau)$  is found as

$$H(t_1 - \tau) = M\Phi(t_1 - \tau)D = \begin{bmatrix} \frac{(t_1 - \tau)^2}{2} & t_1 - \tau \\ t_1 - \tau & 1 \\ 1 & 0 \end{bmatrix} \quad (68)$$

so the row vectors  $h_1(t_1 - \tau)$ ,  $h_2(t_1 - \tau)$ ,  $h_3(t_1 - \tau)$  are

$$\begin{aligned} h_1(t_1 - \tau) &= \left[ \frac{(t_1 - \tau)^2}{2}, t_1 - \tau \right] \\ h_2(t_1 - \tau) &= [t_1 - \tau, 1] \\ h_3(t_1 - \tau) &= [1, 0] \end{aligned} \quad (69)$$

The row vector  $k(t_1 - \tau)$  obtained from eqn (43) is

$$\begin{aligned} k(t_1 - \tau) &= \sum_{j=1}^3 \lambda_j h_j(t_1 - \tau) \\ &= \left[ \frac{\lambda_1}{2}(t_1 - \tau)^2 + \lambda_2(t_1 - \tau) + \lambda_3, \lambda_1(t_1 - \tau) + \lambda_2 \right] \end{aligned}$$

Now since  $p_1 = 2$  and  $p_2 = \infty$ , there is  $q_1 = 2$  and  $q_2 = 1$  which gives

$$\begin{aligned} \|k_1\|_{q_1} &= \left[ \int_{t_0}^{t_1} \left| \frac{\lambda_1}{2}(t_1 - \tau)^2 + \lambda_2(t_1 - \tau) + \lambda_3 \right|^2 d\tau \right]^{1/2} \\ \|k_2\|_{q_2} &= \left[ \int_{t_0}^{t_1} |\lambda_1(t_1 - \tau) + \lambda_2| d\tau \right] \end{aligned} \quad (70)$$

It is desired to minimize

$$\|k\|_1 = L_1 \|k_1\|_2 + L_2 \|k_2\|_1 \quad (71)$$

subject to the constraint  $\lambda e^d(t_1) = \sum_{j=1}^3 \lambda_j e_j^d(t_1) = 1$ . Saying

that the initial condition  $x_0 = \begin{bmatrix} -1 \\ 0 \\ 0 \end{bmatrix}$  which for  $y^d(t_1) = x^d(t_1) = 0$  gives

$$e^d(t_1) = y^d(t_1) - M\phi(t_1 - t_0)x_0 = \begin{bmatrix} 1 \\ 0 \\ 0 \end{bmatrix} \quad (72)$$

(Note: This particular  $x_0$  gives  $e^d(t_1)$  which does not depend on  $t_1$ ; in general, however,  $e^d(t_1)$  will be a function of  $t_1$ .)

Now, putting  $\eta = t_1 - \tau$ ,  $T = t_1 - t_0$  and using eqns (70) and (72) it is found that  $\lambda_1 = 1$  and the minimization of eqn (71) becomes

$$\min_{\lambda_2, \lambda_3} \left\{ L_1 \left[ \int_0^T \left( \frac{\eta^2}{2} + \lambda_2 \eta + \lambda_3 \right)^2 d\eta \right]^{1/2} + L_2 \int_0^T |\eta + \lambda_2| d\eta \right\} \quad (73)$$

Taking partials with respect to  $\lambda_2$  and  $\lambda_3$  and equating to zero gives

$$\lambda_2^* = -\frac{T}{2} \quad \text{and} \quad \lambda_3^* = \frac{T^2}{12} \quad (74)$$

and therefore

$$\|k^*\|_1 = L_1 \left( \frac{T^5}{720} \right)^{1/2} + L_2 \frac{T^2}{4} \quad (75)$$

For any constraints  $L_1$  and  $L_2$  the minimum time interval  $T_0$  is obtained by plotting  $\|k^*\|_1$  in eqn (75) and finding  $T$  which makes it equal to unity.

The optimum inputs will be

$$u_1(t)_{\text{opt}} = \frac{L_1}{\|k_1^*\|_2} \left[ \frac{1}{2}(t_0 - t)^2 + \frac{T_0}{2}(t_0 - t) + \frac{T_0^2}{12} \right] \quad (76)$$

and

$$u_2(t)_{\text{opt}} = L_2 \operatorname{sgn} \left( t_0 - t + \frac{T_0}{2} \right) \quad (77)$$

for  $t_0 \leq t \leq t_0 + T_0$  where  $\|k_1^*\|_2$  in eqn (76) is a number obtained by using  $\lambda_1 = 1$ , eqn (74) and  $T_0$  in eqn (70).

## Conclusions

The procedure presented in this paper shows how to obtain the optimum time solution for a system with multiple input constraints. The form of the solution is readily obtained, but it should be noted that the actual evaluation of the optimum input requires that one be able to find  $\lambda^*$  which gives the minimum of  $\|k\|_1$ . This in itself may be a difficult numerical problem. The procedures suggested by Kulikowski<sup>2, 3</sup> which make use of known minimizing properties of Legendre or Tchebyscheff polynomials and which may be used in the special cases of energy or amplitude constraints respectively, for systems with a single input or with multiple inputs constrained in the same way (i.e.  $p_1 = p_2 = \dots = p_r = p$ ) are not applicable to the minimization of  $\|k\|_1$  as given by eqn (59). When variational methods<sup>10</sup> or the maximum principle of Pontryagin<sup>11</sup> are used to solve optimization problems one finds that a subsidiary mathematical problem also arises. In these cases the subsidiary problem is a two point boundary value problem which is not well suited to a computer solution. On the other hand, when the approach presented in this paper is used, the minimization problem which arises is well adapted to solution by a computer.

## Appendix I, Controllability

The following definition is equivalent to the definition offered by Kalman<sup>4</sup>.

**Definition 1**—A plant is said to be completely controllable if there exists a control input  $u(t)$  which will move the plant

from any initial state  $x(t_0)$  at any initial time  $t_0$  to any desired state in a finite time interval  $[t_0, t_1]$ .

**Definition 2**—A plant is said to be completely output controllable if for any time  $t_0$  and any state  $x(t_0)$  there exists a control input  $u(t)$  which will move the plant initially in  $x(t_0)$  to any desired output in a finite time interval.

It can be shown in a manner identical to that of Kalman<sup>4</sup>, that a necessary and sufficient condition that the system described by eqn (1) be completely output controllable, is that the matrix

$$\Gamma(t_0, t_1) = M(t_1) \left\{ \int_{t_0}^{t_1} \Phi(t_1, t) D(t) D^T(t) \Phi^T(t_1, t) dt \right\} M^T(t_1) \quad (78)$$

be positive definite for some finite time  $t_1 > t_0$ . (The superscript  $T$  denotes matrix transpose.) Furthermore, if the system of eqn (1) is time invariant, then a necessary and sufficient condition for complete output controllability is that

$$\text{rank} \{M[D \ FD \ \dots \ F^{n-1}D]\} = m \quad (79)$$

Note that when condition (78) or (79) is satisfied for  $M = I$  (the identity matrix) then the system is completely controllable. Also note that complete controllability implies but is not implied by complete output controllability. Plants considered in this paper are required to satisfy the weaker condition of complete output controllability.

## Appendix II, A Generalization of Hölder's Inequality

The purpose of this appendix is to obtain an extension of Hölder's inequality to an integral of the form

$$\int_a^b x(t) \cdot y(t) dt = \int_a^b \sum_{i=1}^n x_i(t) y_i(t) dt \quad (80)$$

In the following it is shown that

$$\left| \int_a^b x(t) \cdot y(t) dt \right| \leq \left[ \sum_{i=1}^n L_i^{-p} \|x_i\|_{p_i}^p \right]^{1/p} \left[ \sum_{i=1}^n L_i^q \|y_i\|_{q_i}^q \right]^{1/q} \quad (81)$$

for  $p \geq 1$ ,  $q \geq 1$  and  $1/p + 1/q = 1$  where  $|x_i(t)|^{p_i}$  and  $|y_i(t)|^{q_i}$  are integrable, the  $L_i$  are positive constants, and where

$$\|x_i\|_{p_i} = \left[ \int_a^b |x_i(t)|^{p_i} dt \right]^{1/p_i} \quad (82)$$

$$\|y_i\|_{q_i} = \left[ \int_a^b |y_i(t)|^{q_i} dt \right]^{1/q_i}$$

for  $p_i \geq 1$ ,  $q_i \geq 1$  and  $1/p_i + 1/q_i = 1$ . It is also shown that the inequality in eqn (81) becomes an equality if, and only if

$$x_i(t) = K L_i^q \|y_i\|_{q_i}^{(q-q_i)} |y_i(t)|^{q_i-1} \text{sgn } y_i(t) \quad (83)$$

for all  $a \leq t \leq b$  and all  $i = 1, 2, \dots, n$  where  $K$  is an arbitrary constant.

From the usual form of the Hölder inequality as applied to integrals<sup>5</sup> it is known that

$$\left| \int_a^b x(t) y(t) dt \right| \leq \left[ \int_a^b |x(t)|^p dt \right]^{1/p} \left[ \int_a^b |y(t)|^q dt \right]^{1/q} \quad (84)$$

for  $p > 1$ ,  $q > 1$ , and  $1/p + 1/q = 1$  when  $|x(t)|^p$  and  $|y(t)|^q$  are integrable, and that this inequality becomes an equality if and only if

$$x(t) = K |y(t)|^{q-1} \text{sgn } y(t) \quad (85)$$

where  $K$  is an arbitrary constant. These results can also be shown to hold<sup>1, 9</sup> when  $p = 1$  or  $q = 1$  if condition (85) is properly interpreted.

The equivalent Hölder inequality for sums<sup>5</sup> is

$$\left| \sum_{i=1}^n \alpha_i \beta_i \right| \leq \left[ \sum_{i=1}^n \alpha_i^p \right]^{1/p} \left[ \sum_{i=1}^n \beta_i^q \right]^{1/q} \quad (86)$$

where the inequality holds if and only if

$$\alpha_i = K |\beta_i|^{q-1} \text{sgn } \beta_i \quad \text{for } i = 1, 2, \dots, n \quad (87)$$

where  $K$  is an arbitrary constant.

Again, these results are usually shown to hold for  $1/p + 1/q = 1$ ,  $p > 1$  and  $q > 1$ , but by proper interpretation of eqn (87) they can be shown to hold also for  $p = 1$  or  $q = 1$ .<sup>9</sup>

To obtain a Hölder inequality for the integral of eqn (80) observe that

$$\left| \int_a^b \sum_{i=1}^n x_i(t) y_i(t) dt \right| \leq \sum_{i=1}^n \left| \int_a^b x_i(t) y_i(t) dt \right| \quad (88)$$

and that the equality in eqn (88) holds if and only if

$$x_i(t) y_i(t) \geq 0 \quad [\text{or } x_i(t) y_i(t) \leq 0] \quad (89)$$

for  $a \leq t \leq b$  and all  $i = 1, 2, \dots, n$ .

Now applying the results of the simple Hölder inequality as expressed in eqn (84) it is found that

$$\left| \int_a^b x_i(t) y_i(t) dt \right| \leq \|x_i\|_{p_i} \|y_i\|_{q_i} \quad (90)$$

for  $p_i \geq 1$ ,  $q_i \geq 1$  and  $1/p_i + 1/q_i = 1$  when  $|x_i(t)|^{p_i}$  and  $|y_i(t)|^{q_i}$  are integrable, where  $\|x_i\|_{p_i}$  and  $\|y_i\|_{q_i}$  are defined by eqn (82). By eqn (85) it is found that the inequality in eqn (90) becomes an equality if and only if

$$x_i(t) = K_i |y_i(t)|^{q_i-1} \text{sgn } y_i(t) \quad \text{for } a \leq t \leq b \quad (91)$$

Inserting eqns (80) and (90) into eqn (88) gives

$$\left| \int_a^b x(t) \cdot y(t) dt \right| \leq \sum_{i=1}^n \|x_i\|_{p_i} \|y_i\|_{q_i} \quad (92)$$

Note that if eqn (91) holds for all  $i = 1, 2, \dots, n$  and if all  $K_i$  have the same sign then conditions (89) are satisfied. This means that the equality in eqn (92) is valid if and only if eqn (91) holds for all  $i = 1, 2, \dots, n$  and all  $K_i$  have the same sign.

Now if

$$\hat{x}_i = \frac{\|x_i\|_{p_i}}{L_i} \quad (93)$$

and

$$\hat{y}_i = L_i \|y_i\|_{q_i}$$

where the  $L_i$  are positive constants, then since  $\hat{x}_i$  and  $\hat{y}_i$  are positive it is found that

$$\sum_{i=1}^n \hat{x}_i \hat{y}_i = \left| \sum_{i=1}^n \hat{x}_i \hat{y}_i \right| \quad (94)$$



and using the results of eqns (86) and (87) it is found that

$$\left| \sum_{i=1}^n \hat{x}_i \hat{y}_i \right| \leq \left[ \sum_{i=1}^n |\hat{x}_i|^p \right]^{1/p} \left[ \sum_{i=1}^n |\hat{y}_i|^q \right]^{1/q} \quad (95)$$

where  $p \geq 1$ ,  $q \geq 1$  and  $1/p + 1/q = 1$ , and that the equality holds if and only if

$$\hat{x}_i = \hat{K} |\hat{y}_i|^{(q-1)} \text{sgn } \hat{y}_i \quad \text{for all } i \quad (96)$$

where  $\hat{K}$  is an arbitrary positive constant. However, since  $\hat{x}_i$  and  $\hat{y}_i$  are positive, eqn (96) can be expressed as

$$\hat{x}_i = \hat{K} \hat{y}_i^{(q-1)} \quad (97)$$

where  $\hat{K}$  is a positive constant. When eqn (93) is inserted in eqn (95) and the resulting equation combined with eqn (92) gives

$$\left| \int_a^b x(t) \cdot y(t) dt \right| \leq \left[ \sum_{i=1}^n L_i^{-p} \|x_i\|_{p_i}^p \right]^{1/p} \left[ \sum_{i=1}^n L_i^q \|y_i\|_{q_i}^q \right]^{1/q}$$

which is the desired inequality (81). The equality in eqn (81) is satisfied if and only if the conditions for the equality in eqns (92) and (95) hold. This requires that

- (a) all  $K_i$  have the same sign; (98)
- (b) eqns (91) and (97) hold.

Considering condition (97) it is found that

$$\begin{aligned} \hat{x}_i &\triangleq L_i^{-1} \|x_i\|_{p_i} \triangleq L_i^{-1} \left[ \int_a^b |x_i(t)|^{p_i} dt \right]^{1/p_i} \\ &= \hat{K} \hat{y}_i^{(q-1)} \triangleq \hat{K} L_i^{(q-1)} \|y_i\|_{q_i}^{q-1} \end{aligned} \quad (99)$$

Now insert  $x_i(t)$  as given by eqn (91) in the integral of eqn (99). This gives

$$L_i^{-1} |K_i| \|y_i\|_{q_i}^{q/p_i} = \hat{K} L_i^{(q-1)} \|y_i\|_{q_i}^{q-1} \quad (100)$$

Solving for  $|K_i|$  gives

$$|K_i| = \hat{K} L_i^q \|y_i\|_{q_i}^{q-q_i} \quad \text{for all } i=1, 2, \dots, n \quad (101)$$

Therefore if  $K_i$  in eqn (91) is always chosen so that eqn (101) holds then eqn (97) will also be satisfied. It should also be noted that  $K_i$  may be positive or negative provided that all  $K_i$  have the same sign. This means that eqns (101) and (98) can be replaced by

$$K_i = K L_i^q \|y_i\|_{q_i}^{q-q_i} \quad \text{for all } i=1, 2, \dots, n \quad (102)$$

where  $K$  is an arbitrary constant.

Substituting this result into eqn (91) gives the final form of the necessary and sufficient conditions for an equality in eqn (81) as

$$x_i(t) = K L_i^q \|y_i\|_{q_i}^{q-q_i} |y_i(t)|^{q_i-1} \text{sgn } y_i(t)$$

for  $a \leq t \leq b$  and all  $i = 1, 2, \dots, n$ . This is the desired result eqn (83).

### Appendix III

In order to show that the input  $u(\tau)$  with components given by

$$u_i(\tau) = \frac{L_i^q}{(\|k^*\|_q)^q} \|k_i^*\|_{q_i}^{q-q_i} |k_i^*(t_1, \tau)|^{q_i-1} \text{sgn } k_i^*(t_1, \tau) \quad (103)$$

for  $t_0 \leq \tau \leq t_1$

satisfies the terminal conditions

$$e_j^d(t_1) = \int_{t_0}^{t_1} h_j(t_1, \tau) u(\tau) d\tau \quad (104)$$

for  $j = 1, 2, \dots, m$ , consider the problem of finding

$$\max_{\lambda} \frac{\left| \sum_{j=1}^m \lambda_j e_j^d(t_1) \right|}{\|k\|_q} \quad (105)$$

To do this take

$$\begin{aligned} \frac{\partial}{\partial \lambda_l} \left[ \frac{\left| \sum_{j=1}^m \lambda_j e_j^d(t_1) \right|}{\|k\|_q} \right] \\ = \frac{\|k\|_q \text{sgn} \left[ \sum_{j=1}^m \lambda_j e_j^d(t_1) \right] e_l^d(t_1) - \left| \sum_{j=1}^m \lambda_j e_j^d(t_1) \right| \frac{\partial}{\partial \lambda_l} \|k\|_q}{\|k\|_q^2} = 0 \end{aligned} \quad \text{for } l=1, 2, \dots, m \quad (106)$$

which gives

$$e_l^d(t_1) = \frac{\sum_{j=1}^m \lambda_j e_j^d(t_1)}{\|k\|_q} \frac{\partial}{\partial \lambda_l} [\|k\|_q] \quad (107)$$

From the definition of  $\|k\|_q$  in eqn (46) it is found

$$\frac{\partial}{\partial \lambda_l} \|k\|_q = \frac{1}{q} [\|k\|_q]^{1-q} \sum_{i=1}^r L_i^q \frac{\partial}{\partial \lambda_l} [\|k_i\|_{q_i}^q] \quad (108)$$

and in turn

$$\frac{\partial}{\partial \lambda_l} [\|k_i\|_{q_i}^q] = q \|k_i\|_{q_i}^{q-1} \frac{\partial}{\partial \lambda_l} \|k_i\|_{q_i} \quad (109)$$

From the definition of  $\|k_i\|_{q_i}$  (eqn 45)

$$\begin{aligned} \frac{\partial}{\partial \lambda_l} [\|k_i\|_{q_i}] \\ = [\|k_i\|_{q_i}]^{1-q_i} \int_{t_0}^{t_1} |k_i(t_1, \tau)|^{q_i-1} \text{sgn } k_i(t_1, \tau) \frac{\partial k_i(t_1, \tau)}{\partial \lambda_l} d\tau \end{aligned} \quad (110)$$

From eqn (43) it is found that  $k_i(t_1, \tau)$  is given by

$$k_i(t_1, \tau) = \sum_{j=1}^m \lambda_j h_{ij}(t_1, \tau) \quad (111)$$

where  $h_{ij}(t_1, \tau)$  is the element in the  $i$ th row and  $j$ th column of the matrix  $H(t_1, \tau)$ . This equation shows that

$$\frac{\partial k_i(t_1, \tau)}{\partial \lambda_l} = h_{li}(t_1, \tau) \quad (112)$$

Substituting eqn (112) into (110), eqn (110) into (109), eqn (109) into (108) and eqn (108) into (107) gives

$$\begin{aligned} e_l^d(t_1) &= \frac{\sum_{j=1}^m \lambda_j e_j^d(t_1)}{\|k\|_q} \|k\|_q^{1-q} \\ &= \sum_{i=1}^r L_i^q \|k_i\|_{q_i}^{q-q_i} \int_{t_0}^{t_1} |k_i(t_1, \tau)|^{q_i-1} \text{sgn } k_i(t_1, \tau) h_{li}(t_1, \tau) d\tau \end{aligned} \quad (113)$$

This equation can be rewritten as

$$e_l^d(t_1) = \int_{t_0}^{t_1} \sum_{i=1}^r h_{il}(t_1, \tau) \left\{ \frac{\sum_{j=1}^m \lambda_j e_j^d(t_1)}{\|k\|_q^q} L_i^q \|k_i\|_{dt}^{q-q_i} |k_i(t_1, \tau)|^{q_i-1} \operatorname{sgn} k_i(t_1, \tau) \right\} d\tau$$

for  $l=1, 2, \dots, m$  (114)

Now when the set of eqns (114) are solved for  $\lambda_j$ 's these values of  $\lambda_j$  will give the maximum in eqn (105).

Alternatively if a set of  $\lambda_j$ 's which maximize the expression in eqn (105) are inserted into eqn (114) then these equations are satisfied identically. Recall that the values  $\lambda_j^*$  defined after eqn (48) makes (105) a maximum. Therefore when  $\lambda_j^*$ 's are used in eqn (114) and the quantity in the brackets is denoted as  $u_i(\tau)$  this is precisely the form of  $u_i(\tau)$  given in eqn (103). It can be concluded therefore that when  $u_i(\tau)$  is chosen as in eqn (103) the set of eqns (114) become

$$e_l^d(t_1) = \int_{t_0}^{t_1} \sum_{i=1}^r h_{il}(t_1, \tau) u_i(\tau) d\tau = \int_{t_0}^{t_1} h_l(t_1, \tau) u(\tau) d\tau$$

for  $l=1, 2, \dots, m$  (115)

and are satisfied identically. These equations are seen to be the terminal conditions (104).

## References

- <sup>1</sup> KRANC, G. M. and SARACHIK, P. E. An application of functional analysis to the optimal control problem. *Trans. Amer. Soc. mech. Engrs, J. Basic Engng* 85 (1963) 143
- <sup>2</sup> KULIKOWSKI, R. On optimal control with constraints. *Bull. Polish Acad. Sci. (Ser. tech. Sci.)* 7 (1959) 285
- <sup>3</sup> KULIKOWSKI, R. Optimizing processes and synthesis of optimizing automatic control systems. *Automatic and Remote Control*, p. 473. 1961. London; Butterworths
- <sup>4</sup> KALMAN, R. E. Contributions to the theory of optimal control. *Bol. Soc. Matem. Mex.* (1960)
- <sup>5</sup> LIUSTERNIK, L. and SOBOLEV, V. *Elements of Functional Analysis*. 1962. New York; Ungar
- <sup>6</sup> KIRILLOVA, F. M. A limiting process in the solution of an optimal control process. *Appl. Mech. and Math.* 24 (1960) 398
- <sup>7</sup> KRASOVSKII, N. N. On the theory of optimum regulation. *Automation and Remote Control* 18 (1957) 1005
- <sup>8</sup> CODDINGTON, E. A. and LEVENSON, N. *Theory of Differential Equations*. 1955. New York; McGraw-Hill
- <sup>9</sup> KREINDLER, E. Contributions to the Theory of Time Optimal Control; *J. Franklin Inst.* 275 (1963) 314
- <sup>10</sup> FRIEDLAND, B. The structure of optimal control systems. *Trans. Amer. Soc. mech. Engrs, J. Basic Engng* 84 (1962)
- <sup>11</sup> PONTRYAGIN, L. S. Optimal control processes. *Automat. Express* 1 (1960) 15, 26

# The Approximate Calculation of a Class of Automatic Systems with Forced Parameter Optimization

YU. I. ALIMOV

## Summary

The paper considers: (1) two linear filters  $\Phi_1$  and  $\Phi_2$  connected in parallel, with an input signal  $\theta(t)$ , and (2) closed-loop systems for adjusting the parameters  $X_i$  of filter  $\Phi_2$ , which employ a search modulation  $\delta x_i(t)$  of these parameters. The self-adjustment network includes a detector (linear or square-law) for the output quantity of the filter  $\Phi_2 - \Phi_1$ , phase discriminators and averaging filters  $W_\phi$ .

On the assumption that the amplitudes of the signals  $\delta x_i(t)$  are small, approximate (and, within the limits of the small-parameter method, entirely correct) expressions are found for the control actions in the self-adjustment networks in terms of the frequency characteristics of the filters  $\Phi_1$  and  $\Phi_2$  and the present spectra of the signals  $\delta x_i(t)$  and  $\theta(t)$ . The differential equations derived for the self-adjustment processes take account of the limited memory and pass-band of the filters  $\Phi_1$ ,  $\Phi_2$  and  $W_\phi$ , and cover the case where the frequency characteristic of the filter  $\Phi_2$  is a functional of the signals  $\delta x_i(t)$ . These equations lead readily to a number of necessary conditions for the stability of the self-adjustment process, and also to the desirability, with a high-frequency sinusoidal search signal  $\delta x_i(t)$ , of using two phase discriminators with reference voltages in quadrature, so as to make use also of the phase modulation on the carrier signal.

As a simplified mathematical abstraction, detailed consideration is given to the case of an almost periodic signal  $\theta(t)$  and a test signal similar in nature to white noise. Some attention is devoted also to quasi-stationary self-adjustment modes of operation. Illustrative examples are given of the calculation of actual systems with several adjustable parameters.

## Sommaire

La communication s'attache à l'ensemble de: (1) deux filtres linéaires  $\Phi_1$  et  $\Phi_2$ , connectés en parallèle, recevant un signal d'entrée  $\theta(t)$  et (2) des systèmes en boucle fermée destinés à adapter les paramètres  $X_i$  du filtre  $\Phi_2$  à l'aide d'une modulation exploratoire  $\delta x_i(t)$  de ces paramètres. Le circuit auto-adaptatif comprend un détecteur (linéaire ou quadratique) de la grandeur de sortie du filtre  $\Phi_2 - \Phi_1$ , des discriminateurs de phase et des filtres de moyenne  $W_\phi$ .

A partir de l'hypothèse que les signaux  $\delta x_i(t)$  sont petits, on établit des expressions approchées mais qui, dans les limites de la méthode du petit paramètre, sont parfaitement exactes, déterminant la grandeur des actions de réglage dans les circuits d'auto-adaptation en fonction des caractéristiques de fréquence des filtres  $\Phi_1$  et  $\Phi_2$  et des spectres des signaux  $\delta x_i(t)$  et  $\theta(t)$ . Les équations différentielles obtenues pour les processus auto-adaptatifs tiennent compte du fait que la capacité de mémorisation et la bande passante des filtres  $\Phi_1$  et  $\Phi_2$  et  $W_\phi$  sont limitées; elles s'appliquent dans le cas où la caractéristique de fréquence du filtre  $\Phi_2$  est une fonctionnelle des signaux  $\delta x_i(t)$ . Ces équations conduisent directement à un certain nombre de conditions nécessaires pour que le système auto-adaptatif soit stable. Elles mettent en évidence l'intérêt qu'il y aurait, avec un signal exploratoire  $\delta x_i(t)$  sinusoïdal à haute fréquence, à recourir à des discriminateurs de phase biphasés, à tensions de référence en quadrature, ce qui permettrait d'utiliser également la modulation de phase de la fréquence porteuse.

A titre d'abstraction mathématique simplifiée, on étudie en détail le cas où le signal  $\theta(t)$  est presque périodique et où le signal explora-

toire a la constitution d'un bruit blanc. On s'attache aussi quelque peu aux modes d'auto-adaptation quasi-stationnaires. Des exemples illustrent le calcul de systèmes existants avec plusieurs paramètres réglables.

## Zusammenfassung

Der Aufsatz betrachtet ein Regelsystem das 1. aus zwei parallelen linearen Filtern  $\Phi_1$  und  $\Phi_2$  mit einem Eingangssignal  $\theta(t)$  und 2. einem geschlossenen Kreis zur Einstellung der Parameter  $X_i$  des Filters  $\Phi_2$  besteht, welche für die Selbsteinstellung eine Modulation  $\delta x_i(t)$  dieser Parameter verwendet. Das selbsteinstellende Netzwerk enthält einen linearen oder quadratischen Detektor für das Signal am Filterausgang ( $\Phi_2 - \Phi_1$ ), Phasendiskriminatoren und Glättungsfilter  $W_\phi$ .

Bei der Annahme, daß die Amplituden der Signale  $\delta x_i(t)$  klein sind, findet man für den Regelvorgang der selbsteinstellenden Netzwerke Näherungsausdrücke, die den Frequenzgang der Filter  $\Phi_1$  und  $\Phi_2$  und die vorliegenden Spektren der Signale  $\delta x_i(t)$  und  $\theta(t)$  enthalten (innerhalb der Grenzen der Methode der kleinen Parameter sind diese Ausdrücke völlig exakt). Die für die Selbsteinstellung abgeleiteten Differentialgleichungen berücksichtigen die begrenzte Speicherefähigkeit und die Bandbreite der Filter  $\Phi_1$ ,  $\Phi_2$  und  $W_\phi$  und gelten auch für den Fall, daß der Frequenzgang des Filters  $\Phi_2$  eine Funktion des Signales  $\delta x_i(t)$  ist. Diese Gleichungen führen bereits zu einer Anzahl notwendiger Bedingungen für die Stabilität der Selbsteinstellung. Sie zeigen auch, daß bei sinusförmigen hochfrequenten Signalen zur Selbsteinstellung  $\delta x_i(t)$  die Verwendung zweier Phasendiskriminatoren mit quadratischer Richtkennlinie wünschenswert ist, um auch die Phasenmodulation des Trägersignals auszunutzen.

Zur vereinfachten mathematischen Darstellung wird der Fall eines fastperiodischen Signals  $\theta(t)$  und eines Testsignales, das weißem Rauschen ähnelt, genauer betrachtet. Ein weiterer Abschnitt befaßt sich mit der Selbsteinstellung bei quasi-stationären Vorgängen. Die Arbeit enthält durchgerechnete Beispiele ausgeführter Systeme mit mehreren einstellbaren Parametern.

## Introduction

§ 1. This paper considers an automatic system (hereafter called System A) that consists of linear continuous filters  $\Phi_1$  and  $\Phi_2$  connected in parallel, with a test signal  $\theta(t)$  at the input and closed-loop astatic systems for adjusting the parameters  $X = (X_1, \dots, X_N)$  of filter  $\Phi_2$  (see Figure 1). The self-adjusting circuit includes a detector  $\mathcal{D}$  of the error signal  $\varepsilon(t)$ , phase discriminators  $\Phi_{\mathcal{D}_i}$ , averaging filters  $W_{\phi_i}$  and integrating networks. The control actions in the parameter-adjusting circuits are formed by using a search modulation  $\mu \Delta x(t)$  of the parameters. The defined parameters  $Y_1$  and  $Y_2$  of filters  $\Phi_1$  and  $\Phi_2$  respectively vary with time according to a law that is only known approximately beforehand.

In practice, the following variants of System A are most often met:

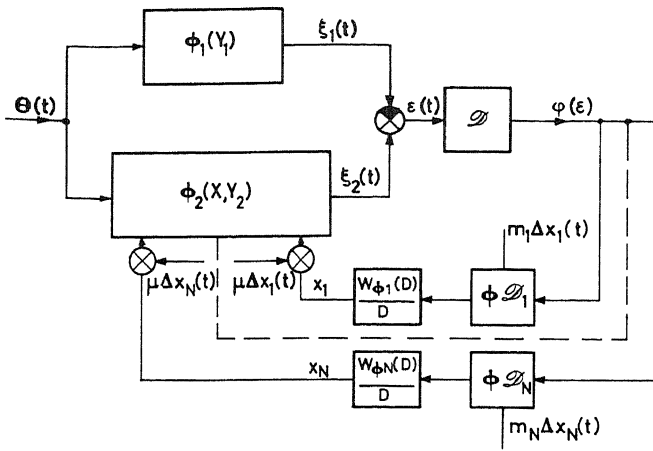


Figure 1

(1)  $Y_1 = \text{const.}$ ,  $Y_2 = Y_2(t)$ . The filter  $\Phi_1$  is a stationary calibration display unit, while the filter  $\Phi_2$  is an automatic system with extremal adjustment of its correcting elements, compensating to a given extent the drift of the parameters  $Y_2(t)^{1-4}$  or the variation in the form of the external action  $\theta(t)^5$ .

(2)  $Y_1 = Y_1(t)$ ,  $Y_2 = \text{const.}$  Filter  $\Phi_1$  is a controlled plant with variable dynamic properties, while filter  $\Phi_2$  is a learning model of this plant<sup>6</sup>.

Of course, the general case  $Y_1 = Y_1(t)$ ,  $Y_2 = Y_2(t)$  is also possible in practice; for example, a calibration display unit  $\Phi_1$  with programmed parameter variation.

§ 2. In Part I of the paper the small-parameter method is used in deriving enough general approximate equations for the processes of self-adjustment in System A under the assumption that the amplitudes of the search signals  $\mu \Delta x(t)$  are small. The equations take account of the limited memory of filters  $\Phi_1$  and  $\Phi_2$ , and cover the case of any given explicit test and search actions. The control signals in the self-adjusting circuits are expressed in terms of the frequency characteristics of filters  $\Phi_1$  and  $\Phi_2$  and the spectra of signals  $\theta(t)$  and  $\mu \Delta x(t)$ . In Part II the general equations of motion for System A are simplified, taking the assumption that the search signals  $\mu \Delta x(t)$  are sinusoidal. Then, as a simplified mathematical abstraction, the case of an almost periodic action  $\theta(t)$  is examined in detail in Part III. A very simple analysis of the relevant equations of motion shows the desirability, with a high-frequency sinusoidal signal  $\mu \Delta x_i(t)$ , of using, in the phase discriminator, a reference voltage phase shifted with respect to this signal, which permits one to make use of the extra useful information carried by the quadrature component of the search-frequency signal, by analogy with the practice, in radio engineering, of using amplitude and phase modulation simultaneously<sup>7</sup>. Part IV uses the example of a white-noise test signal to show that the equations derived may also be applied to the description of System A with stochastically defined signals  $\theta(t)$ , without relying on the hypothesis of the closeness of random processes in the system to stationary ergodic ones. There is a brief discussion of the relation between the results derived here and those in previous papers<sup>1-6</sup>. Some attention is also devoted to quasi-stationary modes of self-adjusting operation.

In conclusion it should be stressed that all the design examples

quoted have been chosen to be simple as far as possible, and that the main emphasis is on the physical interpretation and qualitative analysis of the mathematical relations derived.

### I. Derivation of General Equations of Motion for the Self-adjusting System Considered

§ 3. The most important of the assumptions, under which the equations for the processes of self-adjustment in System A are derived below, are first set out:

(a) The amplitudes of the search-modulation signals are considered small, and to emphasize this they are denoted by  $\mu \Delta x(t)$  where  $\mu$  is a small parameter.

(b) It is assumed that System A starts to operate at a certain instant  $t = t_0$ , having been in an equilibrium condition up to that time, and thus the output quantity of filter  $\Phi_i$  ( $i = 1, 2$ ) is determined by the relation

$$\xi_i(t) = \int_{t_0}^t \theta(\tau) K_i(t, \tau) d\tau, \quad i = 1, 2 \quad (1)$$

where  $K_i(t, \tau)$  is the weighting function of filter  $\Phi_i$ .

Equation (1) is expressed in the form

$$\xi_i(t) = \int_{t_0}^{t-T} \theta(\tau) K_i(t, \tau) d\tau + \int_{t-T}^t \theta(\tau) K_i(t, \tau) d\tau, \quad i = 1, 2 \quad (2)$$

$$t_0 < t - T < t$$

Let the filters  $\Phi_i$  be stable. Then if

$$|\theta(t)| < \text{const} \quad (-\infty < t < +\infty) \quad (3)$$

it may be considered that for a certain sufficiently large  $T$  the first integral in eqn (2) is negligibly small, and

$$\xi_i(t) \approx \int_{t-T}^t \theta(\tau) K_i(t, \tau) d\tau, \quad i = 1, 2 \quad (4)$$

In other words, this means that by the instant  $t$  information on the state of filters  $\Phi_i$  and on the values of the signal  $\theta(\tau)$  at instants  $\tau > t - T$  is practically completely lost, and the value of  $\xi_i(t)$  may be identified with the reaction of filter  $\Phi_i$  to the signal

$$\theta_T(\tau) = \begin{cases} \theta(\tau) & \text{for } t - T < \tau \leq t \\ 0 & \text{outside that interval} \end{cases} \quad (5)$$

assuming that  $\xi_i(\tau) \equiv 0$  for  $\tau < t - T$ .

The conditional nature of any choice of a numerical value for  $T$  matches the complexity of the actual situation: the 'memory'  $T$  of a linear system depends substantially on the criterion chosen and on many factors that are often not subject to any sort of accurate quantitative calculation (on the structure of the signal  $\theta(t)$  within the bounds of the natural and easily enough controlled restriction (3), on the level of fluctuating disturbance in the system, etc.). If filter  $\Phi_i$  is near to the stability boundary in parameter space, then of course  $T \rightarrow \infty$ . If stability is lost then (4) is not true even for  $T = \infty$ , and strictly speaking, one cannot apply either the theory developed below, which takes no account of the initial perturbations always existing in a system, or the theories of Krasovski<sup>2</sup>, Kazakov<sup>3</sup> and Varygin<sup>4</sup>.

(c) It is also considered that the variation in parameters  $Y_1(t)$ ,  $Y_2(t)$  and  $X(t)$  over the time interval  $T$  may be neglected.

The time-dependence of the frequency characteristics  $W_1(j\omega) = W_1(j\omega, t)$  and  $W_2(j\omega) = W_2(j\omega, t)$  of filters  $\Phi_1$  and  $\Phi_2$  (with  $\Delta x(t) \equiv 0$ ) is only expressed in the taking of the values of parameters  $Y_1(t)$ ,  $Y_2(t)$  and  $X(t)$  as 'frozen' at the given instant  $t$ :

$$Y_1(\tau) \approx Y_1(t), Y_2(\tau) \approx Y_2(t), X(\tau) \approx X(t) \text{ for } t - T < \tau < T \quad (6)$$

(d) Finally, for the sake of definition, it is assumed that the state of filter  $\phi_2$  is described by the ordinary differential equation

$$\sum_{k=0}^n a_{2,k}(Y_2, X + \mu \Delta x) D^k \xi_i(t) = \sum_{l=0}^m b_{2,l}(Y_2, X + \mu \Delta x) D^l \theta(t) \quad (7)$$

$$D \equiv \frac{d}{dt}, \quad m \leq n$$

with coefficients  $a_{2,k}(Y_2, X)$  and  $b_{2,l}(Y_2, X)$  that are analytic in  $X$ . It is evident that

$$W_2(j\omega) = R_2(j\omega) \cdot Q_2^{-1}(j\omega) \quad (8)$$

where

$$Q_2(D) = \sum_{k=0}^n a_{2,k}(Y_2, X) D^k \quad (9)$$

$$R_2(D) = \sum_{l=0}^m b_{2,l}(Y_2, X) D^l$$

Given assumptions (a), (b) and (c) the proposed method of calculation can be generalized without much complication to the case where pure delays are present in the filter  $\Phi_2$  under adjustment.

§ 4. It is observed that assumption (b) allows one to make calculation in a frequency region bounded only by consideration of the 'shortened' present spectrum<sup>8</sup>

$$\theta_T(j\omega, t) = \int_{t-T}^t \theta(\tau) e^{-j\omega\tau} d\tau = \int_{-\infty}^t \theta_T(\tau) e^{-j\omega\tau} d\tau \quad (10)$$

of the signal  $\theta(t)$ . Thus, in particular, taking into account the quasi-stationary nature of the filter  $W_1(j\omega, t)$  the following relations are obtained for  $\xi_1(t)$ :

$$\xi_1(t) = \frac{1}{2\pi} \int_{-\infty}^{\infty} \xi_1(j\omega, t) \cdot e^{j\omega t} d\omega \quad (11)$$

where

$$\xi_1(j\omega, t) \approx W_1(j\omega, t) \cdot \theta_T(j\omega, t) \quad (12)$$

Considering, instead of normal spectra, the 'shortened' present spectra of the type in (10) and (12), generally one can reflect more accurately in a mathematical model the actual situations that arise in the experimental development of System A, and also simplify mathematical operations on the spectra of the signals  $\theta(t)$  and  $\xi_i(t)$  in those cases where the Fourier integrals for these functions over the interval  $(-\infty, t)$  diverge. This approach turns out, in particular, to be very convenient for the examination of non-ergodic random processes in System A, as it gives a natural transition to the description of the system in terms of spectral power densities (see Part IV).

Since this paper only considers explicit (and, what is more, only harmonic) search signals  $\mu \Delta x(t)$ , from now on in order to simplify the text the 'full' spectrum is used as a convenient, if less accurate, mathematical abstraction

$$\mu \Delta x_i(j\omega) = \mu \int_{-\infty}^{\infty} \Delta x_i(\tau) \cdot e^{-j\omega\tau} d\tau \quad (13)$$

of the search signal.

§ 5. A solution  $\xi_2(t)$  to eqn (7) is looked for in the form of a series

$$\xi_2(t) = \xi_{20}(t) + \mu \cdot \xi_{21}(t) + \dots \quad (14)$$

all the analysis below being taken only with the accuracy of magnitudes of the first order of magnitude with respect to the quantity  $\mu$  [obviously one way of making the theory more accurate is to take account of more terms in (14)]. Using the normal procedure<sup>9</sup> for the small-parameter method, the following equation for sequential calculation of the quantities  $\xi_{20}(t)$  and  $\xi_{21}(t)$  are obtained from (7)–(9):

$$Q_2(D) \xi_{20}(t) = R_2(D) \theta(t), \quad D \equiv \frac{d}{dt} \quad (15)$$

$$Q_2(D) \xi_{21}(t) = \sum_{i=1}^N \Delta x_i(t) \left[ \frac{\partial R_2(D)}{\partial X_i} \theta(t) - \frac{\partial Q_2(D)}{\partial X_i} \xi_{20}(t) \right] \quad (16)$$

It is easily seen that given assumption (b) the memory of the linear system (16) should be considered as limited to the time interval  $T$ . Hence, taking into account the quasi-stationary nature of the filters  $W_1(j\omega, t)$  and  $W_2(j\omega, t)$  and the identity  $\partial W_1 / \partial X_i \equiv 0$ , the following expression is found for the 'shortened' present spectrum of the error

$$\varepsilon(t) = \xi_{20}(t) - \xi_1(t) - \mu \xi_{21}(t) \quad (17)$$

$$\varepsilon(j\omega, t) \approx W(j\omega) \theta_T(j\omega, t) + \frac{\mu}{2\pi} Q_2^{-1}(j\omega) \sum_{i=1}^N \int_{-\infty}^{\infty} \frac{\partial W(j\nu)}{\partial X_i} Q_2(j\nu) \theta_T(j\nu, t) \Delta x_i(j(\omega - \nu)) \cdot d\nu \quad (18)$$

where  $W(j\omega) = W_2(j\omega) - W_1(j\omega)$

while the spectra  $\theta_T(j\omega, t)$  and  $\Delta x_i(j\omega)$  are defined by eqns (10) and (13).

Furthermore, in accordance with the circuit shown in Figure 1, one obtains for the  $X_i$ -adjusting network

$$DX_i = W_{\phi i}(D) \varphi\{\varepsilon(t)\} \Delta x_i(t), \quad D \equiv \frac{d}{dt} \quad (19)$$

where  $\varphi(\varepsilon)$  is the detector characteristic, while

$$\varepsilon(t) = \frac{1}{2\pi} \int_{-\infty}^{\infty} \varepsilon(j\omega, t) \cdot e^{j\omega t} d\omega \quad (20)$$

The approximate system of eqns (17)–(20) that has been obtained describes a very wide class of self-adjusting operating conditions for System A.

The following points are stressed:

(1) These equations, written in terms of the frequency characteristics, are differential equations (generally speaking, non-linear) and in a number of cases are capable of more effective investigation than the integro-differential equations derived by Krasovski<sup>2</sup> and Varygin<sup>4</sup> in terms of weighting functions.

(2) In distinction to the previous papers quoted<sup>2-4</sup>, the derivation of eqns (17)–(20) does not rely on the assumption that the weighting function, and consequently also the transfer function, of filter  $\Phi_2$  is actually a function rather than a functional of the signals  $\mu\Delta x(t)$ .

## II. Simplification of the General Self-adjustment Equations for the Case of a Harmonic Search Modulation and a Square-law Detector

§ 6. If during the whole time of operation of System A the search  $\mu\Delta x(t)$  is nearly harmonic, then it is convenient to consider that

$$\Delta x_i(t) = \Delta_i \cos \Omega_i t \text{ for } -\infty < t < \infty, \quad i = 1, \dots, N \quad (21)$$

$$(\Omega_i < \Omega_{i+1})$$

Then in accordance with (13)

$$\Delta x_i(j\omega) = \pi [\delta(\omega + \Omega_i) + \delta(\omega - \Omega_i)] \quad (22)$$

Substituting (22) in (17) and using the known properties of  $\delta$  functions, one finds:

$$\begin{aligned} \varepsilon(j\omega, t) = & W(j\omega) \theta_T(j\omega, t) \\ & + \frac{1}{2} \mu Q_2^{-1}(j\omega) \times \\ & \sum_{i=1}^N \Delta_i \left[ \frac{\partial W \{j(\omega + \Omega_i)\}}{\partial X_i} Q_2 \{j(\omega + \Omega_i)\} \theta_T \{j(\omega + \Omega_i), t\} \right. \\ & \left. + \frac{\partial W \{j(\omega - \Omega_i)\}}{\partial X_i} Q_2 \{j(\omega - \Omega_i)\} \theta_T \{j(\omega - \Omega_i), t\} \right] \quad (23) \end{aligned}$$

Then let

$$\begin{aligned} W(j\omega) &= |W(j\omega)| e^{j\varphi} (\varphi = \varphi(\omega)), \\ \theta_T(j\omega, t) &= |\theta_T(j\omega, t)| e^{j\alpha} \quad (\alpha = \alpha(\omega)) \quad (24) \end{aligned}$$

Taking into account the even nature of the amplitude spectrum and the odd nature of the phase spectrum in (24), one can readily deduce from (20) and (22) the following expression for the error  $\varepsilon(t)$ :

$$\begin{aligned} \varepsilon(t) \approx & \frac{1}{\pi} \int_0^\infty \left\{ |W(j\omega)| \cos(\omega t + \varphi + \alpha) \right. \\ & + \frac{1}{2} \mu \sum_{i=1}^N \Delta_i [ \operatorname{Re} C_i \cos(\omega t + \alpha) \\ & \left. - \operatorname{Im} C_i \sin(\omega t + \alpha) ] |\theta_T(j\omega, t)| \right\} d\omega \quad (25) \end{aligned}$$

$$C_i = \frac{\partial W(j\omega)}{\partial X_i} Q_2(j\omega) \left[ \frac{e^{-j\Omega_i t}}{Q_2 \{j(\omega - \Omega_i)\}} + \frac{e^{-j\Omega_i t}}{Q_2 \{j(\omega + \Omega_i)\}} \right] \quad (26)$$

In calculating the passage of the signal  $\varepsilon(t)$  through the detector  $\mathcal{D}$ , it is convenient first to separate, in each term of the integrand in (24) that is enclosed in square brackets, the components in phase with the signal  $\cos(\omega t + \varphi + \alpha)$  and those in quadrature with it. One obtains as a result:

$$\begin{aligned} \varepsilon(t) \approx & \pi^{-1} \int_0^\infty \left\{ \left[ |W(j\omega)| + \frac{1}{2} \mu \sum_{i=1}^N \Delta_i A_i(\omega, t) \right] \right. \\ & \cos(\omega t + \varphi + \alpha) + \frac{1}{2} \mu \sum_{i=1}^N \Delta_i B_i(\omega, t) \\ & \left. \sin(\omega t + \varphi + \alpha) \right\} |\theta_T(j\omega, t)| d\omega \quad (27) \end{aligned}$$

where

$$\begin{aligned} A_i(\omega, t) &= a_i(\omega) \cos \Omega_i t + b_i(\omega) \sin \Omega_i t \\ a_i(\omega) &= \frac{\partial |W(j\omega)|}{\partial X_i} (M_i^+(\omega) \cos \varphi_i^+(\omega) + M_i^-(\omega) \cos \varphi_i^-(\omega)) \\ & - |W(j\omega)| \frac{\partial \varphi}{\partial X_i} (M_i^+(\omega) \sin \varphi_i^+(\omega) + M_i^-(\omega) \sin \varphi_i^-(\omega)) \\ b_i(\omega) &= \frac{\partial |W(j\omega)|}{\partial X_i} (M_i^-(\omega) \sin \varphi_i^-(\omega) - M_i^+(\omega) \sin \varphi_i^+(\omega)) \\ & - |W(j\omega)| \frac{\partial \varphi}{\partial X_i} (M_i^+(\omega) \cos \varphi_i^+(\omega) - M_i^-(\omega) \cos \varphi_i^-(\omega)) \quad (28) \end{aligned}$$

$$\begin{aligned} \frac{Q_2(j\omega)}{Q_2 \{j(\omega - \Omega_i)\}} &= M_i^-(\omega) e^{j\varphi_i^-(\omega)} \\ \frac{Q_2(j\omega)}{Q_2 \{j(\omega + \Omega_i)\}} &= M_i^+(\omega) e^{j\varphi_i^+(\omega)} \quad (29) \end{aligned}$$

[the expressions defining the coefficients  $B_i(\omega, t)$  are analogous to formulae (28) and (29)]. In quasi-stationary self-adjustment modes, when  $\Omega_N$  is small compared with the actual frequencies  $\omega$  of the test signal  $\theta(t)$ ,  $Q_2(j\omega) \approx Q_2(j(\omega \pm \Omega_i))$ , so that

$$M_i^- e^{j\varphi_i^-} \approx M_i^+ e^{j\varphi_i^+} \quad (30)$$

$$a_i(\omega) \approx 2 \frac{\partial |W(j\omega)|}{\partial X_i}, \quad b_i(\omega) \approx 0 \quad (31)$$

§ 7. Most often the detector  $\mathcal{D}$  may be considered as either square-law:

$$\varphi(\varepsilon) = \varepsilon^2 \quad (32)$$

or linear:

$$\varphi(\varepsilon) = |\varepsilon| = (\varepsilon^2)^{\frac{1}{2}} \quad (33)$$

In both cases the theoretical analysis requires the square of the error  $\varepsilon(t)$  to be calculated. Taking only terms of zero- and first-order magnitude with respect to quantity  $\mu$ , the following expression is readily derived from (27):

$$\begin{aligned} \varepsilon^2(t) \approx & \frac{1}{2\pi^2} \int_0^\infty \int_0^\infty D(\omega, \nu) [\cos((\omega - \nu)t + \varphi_\omega - \varphi_\nu) \\ & + \cos((\omega + \nu)t + \varphi_\omega + \varphi_\nu)] d\omega d\nu \quad (34) \end{aligned}$$

where

$$\begin{aligned} D(\omega, \nu) \approx & |\theta_T(j\omega, t)| |\theta_T(j\nu, t)| |W(j\omega)| |W(j\nu)| \\ & \times \left[ 1 + \frac{\mu}{2} \sum_{i=1}^N \Delta_i \{ A_i(\omega, t) |W(j\omega)|^{-1} \right. \\ & \left. + A_i(\nu, t) |W(j\nu)|^{-1} \} \right] \quad (35) \end{aligned}$$

$$\text{while } \vartheta_\omega = \varphi(\omega) + \alpha(\omega) + \psi(\omega) \quad (36)$$

$$\tan \psi(\omega) = \frac{\mu}{2} \sum_{i=1}^N \Delta_i B_i(\omega, t) \left[ |W(j\omega)| + \frac{\mu}{2} \sum_{i=1}^N \Delta_i A_i(\omega, t) \right]^{-1}$$

In all the following, harmonic search-modulation signals are, in fact, considered and as a mathematical model of System A one takes the system of eqns (19), (32)–(36). These equations continue to hold adequately until the instant when through the operation of the self-adjustment circuits the relation

$$|W(j\omega)| \approx \mu N \max_{i, t} \Delta_i |A_i(\omega, t)| \quad (37)$$

becomes true (in this event the approximate expressions in (35) for  $\mathcal{D}(\omega, \nu)$  are already invalid).

### III. Theoretical Analysis of Self-adjustment Modes with an Almost Periodic Test Signal

§ 8. If over the whole time interval that this paper is concerned with the test signal may be represented accurately enough in the form

$$\theta(t) = \theta_0 + \sum_{k=1}^M \theta_k \cos(\omega_k t + \alpha_k) \quad (38)$$

$\theta_k, \alpha_k, \omega_k = \text{const}, \quad \omega_k < \omega_{k+1}$

then it is convenient to consider (38) as being true for  $-\infty < t < +\infty$ . Then

$$|\theta_T(j\omega, t)| = \pi \sum_{k=0}^M \theta_k \delta(\omega - \omega_k), \quad \omega_0 = 0 \quad (39)$$

and in accordance with (32)–(35)

$$\varepsilon^2(t) = \frac{1}{2} \sum_{k, l=0}^M \theta_k \theta_l |W(j\omega_k)| |W(j\omega_l)| e_{kl} \quad (40)$$

$$e_{kl} = \left[ 1 + \frac{\mu}{2} \sum_{i=1}^N \Delta_i A_i(\omega_k, t) |W(j\omega_k)|^{-1} + A_i(\omega_l, t) |W(j\omega_l)|^{-1} \right] \times [\cos((\omega_k - \omega_l)t + \vartheta_k - \vartheta_l) + \cos((\omega_k + \omega_l)t + \vartheta_k + \vartheta_l)] \quad (41)$$

where  $\vartheta_k$  and  $A_i(\omega_k, t)$  are as defined by (36) and (28).

It can be seen from (41) and (28) that the signal  $e_{kl}$  is made up of a sum of harmonics at frequencies  $\omega_k \pm \omega_l, \omega_k \pm \omega_l \pm \Omega_s$  ( $k, l = 0, \dots, M, s = 1, \dots, N$ ). In the phase discriminator of the  $i$ th self-adjustment channel the output quantity  $\varphi(\varepsilon)$  of the detector  $\mathcal{D}$  is multiplied by the harmonic reference voltage at frequency  $\Omega_i$ , so that with square-law detection a signal  $u_i(t)$  is obtained consisting of harmonics at frequencies  $\omega_k \pm \omega_l \pm \Omega_i, \omega_k \pm \omega_l \pm \Omega_s \pm \Omega_i$  (with linear detection in general one also gets other harmonic components with amplitudes that are first-order of magnitude with respect to quantity  $\mu$ ).

If in (38)  $\theta_0$  represents a slowly varying useful signal, while the sum

$$\sum_{k=1}^M \theta_k \cos(\omega_k t + \alpha_k) \quad (42)$$

represents intense disturbances at sufficiently high frequencies ( $\omega_1 > 2\Omega_N$ ), then correctly chosen smoothing filters  $W_{\phi i}(D)$  should pass only harmonics of the signal  $u_i(t)$  with frequencies

$$\omega_k - \omega_l - \Omega_i, \quad \omega_k - \omega_l - (\Omega_s \pm \Omega_i), \quad k \geq l \quad (43)$$

It is assumed that the disturbances (42) acting on the system are such that for  $k > l$  the conditions

$$\omega_k - \omega_l \neq \Omega_i, \quad \omega_k - \omega_l \neq (\Omega_s \pm \Omega_i) \quad (44)$$

are satisfied with adequate margin. Then it may be considered that the constant component in the signal  $u_i(t)$  that is passed by the filter  $W_{\phi i}(D)$  accurately matches that harmonic of the sequence  $\omega_k - \omega_l - (\Omega_s - \Omega_i)$  for which  $k = l$  and  $s = i$ . Then according to (41) and the circuit of System A

$$\frac{dX_i}{dt} \approx W_{\phi i}(D) \cdot E_t(\varepsilon^2(t)) m_{ic} \cos \Omega_i t \quad (45)$$

where

$$E_t[f(\tau)] = \lim_{T \rightarrow \infty} \frac{1}{2T} \int_{-T}^T f(\tau) d\tau$$

and the values of the parameters  $X, Y_1$  and  $Y_2$  in the expression for  $\varepsilon^2(t)$  are taken as 'frozen' at the instant  $t$  [see (6)], so that

$$E_t[\varepsilon^2(t) m_{ic} \cos \Omega_i t] = \frac{1}{4} \mu \Delta_i m_{ic} [E_{ci}^{(1)} + E_{ci}^{(2)} + E_{ci}^{(3)}] \quad (46)$$

$$E_{ci}^{(1)} = \frac{\partial}{\partial X} \sum_{k=0}^M \theta_k^2 |W(j\omega_k)|^2 (M_{ik}^+ \cos \varphi_{ik}^+ + M_{ik}^- \cos \varphi_{ik}^-) \quad (47)$$

$$E_{ci}^{(2)} = - \sum_{k=0}^M \theta_k^2 |W(j\omega_k)|^2 \frac{\partial}{\partial X} (M_{ik}^+ \cos \varphi_{ik}^+ + M_{ik}^- \cos \varphi_{ik}^-) \quad (48)$$

$$E_{ci}^{(3)} = - \sum_{k=0}^M \theta_k^2 |W(j\omega_k)|^2 \frac{\partial \varphi(\omega_k)}{\partial X_i} (M_{ik}^+ \sin \varphi_{ik}^+ + M_{ik}^- \sin \varphi_{ik}^-) \quad (49)$$

where the quantities  $M_{ik}^\pm e^{j\varphi_{ik}^\pm}$  are defined by formulae (29) with  $\omega = \omega_k$ .

§ 9. The case of the quasi-stationary mode of operation ( $\theta_0 = 0$  and condition (30) satisfied) are first considered. Equation (45) for the self-adjustment process becomes

$$\frac{dX_i}{dt} \approx W_{\phi i}(D) \cdot \frac{1}{2} \mu \Delta_i m_{ic} \frac{\partial}{\partial X} \sum_{k=1}^M \theta_k^2 |W(j\omega_k)|^2 \quad (50)$$

and thus as a result of the normal operation of the self-adjusting circuit (without loss of stability, without intense distortion caused by disturbances etc.) the quantity

$$\sum_{k=1}^M \theta_k^2 |W_2(j\omega_k) - W_1(j\omega_k)|^2 \quad (51)$$

will be a minimum, i.e. in the complex plane the frequency characteristic of the filter being adjusted will approach that of the calibration filter at the points  $\omega = \omega_k$  ( $k = 1, \dots, M$ ) in some mean-square sense. If by varying the adjusted parameters  $X$  the frequency characteristics  $W_1(j\omega)$  and  $W_2(j\omega)$  can be made practically identical over some range of frequencies, then this approach will merely signify that over the given frequency range  $W_1(j\omega) \approx W_2(j\omega)$ , and the result of the normal operation of the self-adjusting circuit will prove practically independent of the actual spectral composition of the test signal  $\sum_{k=1}^M \theta_k \cos(\omega_k t + \alpha_k)$ , ( $\omega_1 \gg \Omega_N$ ) (see Example 1). The latter statement is not valid (see Taylor<sup>5</sup>, and also Example 2) if the filters

$W_1(j\omega)$  and  $W_2(j\omega)$  essentially cannot be made identical. In this event the closest convergence of the frequency characteristics  $W_1(j\omega)$  and  $W_2(j\omega)$  takes place at those points  $\omega = \omega_k$  corresponding to large amplitudes  $\theta_k$ , and the nature of this convergence will change with variation both of the frequencies  $\omega_k$  and of the ratios between the amplitudes  $\theta_k$ .

**Example 1.** In a System A with square-law detector, let filter  $W_1$  be a controlled plant with transfer function  $W_1(p) = (b_2p + b_1)^{-1}$ , and filter  $W_2$  a self-learning model<sup>6</sup> with a transfer function of the form  $W_2(p) = (X_2p + X_1)^{-1}$ , where  $X_1$  and  $X_2$  are the adjustable parameters, by varying which a complete identity between the dynamic properties of model and plant can, in principle, be achieved. If

$$\theta(t) = \theta \cos(\Omega t + \alpha), \quad \Omega \gg \Omega_2 > \Omega_1 \quad (52)$$

then eqn (50) for the quasi-stationary self-adjustment mode takes the following form:

$$\begin{aligned} \dot{X}_1 &= k_1 W_{\phi 1}(D) \\ &\cdot [(X_1 - b_1)(X_2^2 \Omega^2 + X_1 b_1) - \Omega^2 (X_2 - b_2)^2 X_1] \end{aligned} \quad (53)$$

$$\begin{aligned} \dot{X}_2 &= k_2 W_{\phi 2}(D) \Omega^2 \\ &\cdot [(X_2 - b_2)(X_1^2 + X_2 b_2 \Omega^2) - X_2 (b_1 - x_1)^2] \end{aligned} \quad (54)$$

$$k_i = \mu \Delta_i m_{ic} \theta^2 (b_1^2 + b_2^2 \Omega^2)^{-1} \cdot (X_1^2 + X_2^2 \Omega^2)^{-2}, \quad i = 1, 2 \quad (55)$$

It can be seen from eqns (53)–(55) that as a result of the normal operation of the self-adjusting circuit  $X_1 \rightarrow b_1$  and  $X_2 \rightarrow b_2$ , i.e. in fact  $W_2(j\omega) \rightarrow W_1(j\omega)$ , at whatever frequency  $\Omega$  the test signal (52) is applied. The self-adjustment process forms a coupled control of the parameters  $X_1$  and  $X_2$ . The higher the frequency  $\Omega$  of the test signal, the more intensively the adjustment of  $X_2$  takes place (cf. Margolis and Leonides<sup>6</sup>). The stability for small variations of the equilibrium  $X_1 = b_1$ ,  $X_2 = b_2$  in the non-linear system (53)–(55) can readily be examined from the first-order approximation equations.

**Example 2.** If in the system just considered the controlled plant is close in its dynamic properties to the link  $W_1(p) = e^{-p\tau} (b_2p + b_1)^{-1}$ , while  $W_2(p) = (X_2p + X_1)^{-1}$ , then for  $\tau \neq 0$  complete identity of filters  $W_2$  and  $W_1$  cannot be achieved by self-adjustment. Transcribing eqn (50) for this process, it can easily be seen that the result

$$\begin{aligned} X_1 &\rightarrow b_1 \cos \Omega \tau - b_2 \Omega \sin \Omega \tau, \\ X_2 &\rightarrow b_2 \cos \Omega \tau + b_1 \Omega^{-1} \sin \Omega \tau \end{aligned}$$

of normal operation of the self-adjustment network may already depend substantially on the frequency of the test signal (52).

§ 10. Eqns (45)–(49) also permit a number of conclusions of a qualitative nature on non-quasi-stationary modes of self-adjustment in system A ( $\theta_0 \neq 0$ , conditions (30) not satisfied) to be immediately drawn.

It is first observed that the equation

$$Q_2 = Q_2(j\omega) \quad (56)$$

defines a Mikhaylov hodograph<sup>10</sup> for a stable [assumption (b)]

linear system, and consequently the curve (56) has a form similar to that in Figure 2.

It then becomes clear from (45)–(49) that within the limits of the errors introduced by the terms  $E_{ci}^{(2)}$  and  $E_{ci}^{(3)}$  the normal operation of the  $i$ th self-adjustment channel reduces to the conditional minimization of the quantities

$$\sum_{k=0}^M \theta_k^2 |W(j\omega_k)|^2 (M_{ik}^+ \cos \varphi_{ik}^+ + M_{ik}^- \cos \varphi_{ik}^-) \quad (57)$$

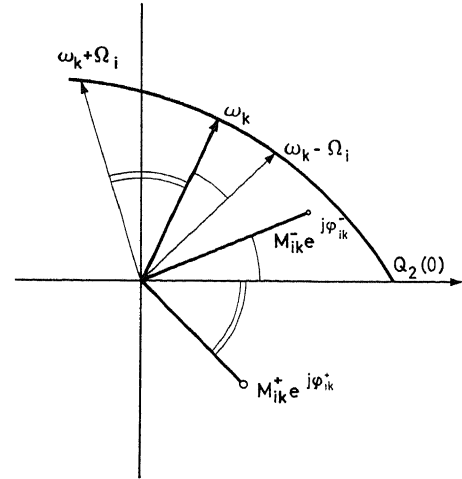


Figure 2

The self-adjustment error associated with  $E_{ci}^{(2)}$  will be small in most cases, since by the very sense of the quantities  $M_{ik}^\pm e^{j\varphi_{ik}^\pm}$  [see (29) and also Figure 2] the partial derivatives  $\partial/\partial X_i (M_{ik}^+ \cos \varphi_{ik}^+ + M_{ik}^- \cos \varphi_{ik}^-)$  will hardly be significantly different from zero. The error associated with  $E_{ci}^{(3)}$  will also be insignificant, since with  $|\varphi_{ik}^\pm|, |\varphi_{ik}^-| < \pi$  the terms  $M_{ik}^+ \sin \varphi_{ik}^+$  and  $M_{ik}^- \sin \varphi_{ik}^-$  are opposite in sign.

If

$$M_{ik}^+ \cos \varphi_{ik}^+ + M_{ik}^- \cos \varphi_{ik}^- > 0, \quad k = 0, \dots, M \quad (58)$$

then the conditional minimization of the quantities (57) has roughly the same physical significance (see § 9) as the minimization of quantity (51) in quasi-stationary modes, and thus the result of normal operation of the self-adjustment network should be taken as acceptable. But the more strongly the self-adjustment mode differs from quasi-stationary, the larger are the angles  $|\varphi_{ik}^\pm|$  for  $|\varphi_{ik}^\pm| < \pi$  the smaller are the coefficients in (58) (in particular, the quantity  $M_{i0}^+ \cos \varphi_{i0}^+ + M_{i0}^- \cos \varphi_{i0}^-$ ), since over a substantial range of frequencies the quantities  $M_{ik}^\pm$  are hardly much different from unity. As a result the quality factor for the  $X_i$  tracking system falls, while for  $|\varphi_{ik}^\pm| > \pi/2$  the coefficient (58) becomes negative and minimization of the weighted sums (57) of squares  $|W_2(j\omega_k) - W_1(j\omega_k)|^2$  loses its evident sense, or even an inversion of the self-adjusting servo-system occurs (particularly if all the coefficients (58) become negative, which may happen if a strong signal  $\theta(t)$  is applied at a frequency close to a search frequency  $\Omega_s$ —see Example 3).

**Example 3.** Let the adjustable filter in System A be a link with transfer function  $W_2(p) = kQ_2(p)^{-1} = k(p^2 + 2\alpha p + \omega_0^2)^{-1}$ , the adjustment parameter being the gain  $k$ , modulated by a signal  $\Delta k \cdot \cos \Omega t$ , while to the input of the system is applied the test action  $\theta(t) = \theta \cos(\omega t + \alpha)$ .



Writing out the general expressions for the quantities

$$M^{\pm} e^{j\varphi^{\pm}} = Q_2(j\omega) \cdot Q_2[j(\omega \pm \Omega)]^{-1} \quad (59)$$

it can readily be established that condition (58) for the system considered is explicitly unobserved if  $\omega_0$  is small ( $\omega_0 \rightarrow 0$ ), while the frequencies  $\Omega$  and  $\omega$  of the search and test signals coincide and exceed  $\omega_0$ , since then

$$M^+ \cos \varphi^+ \rightarrow \frac{1}{4} (\Omega^2 + 2\alpha^2) (\Omega^2 + \alpha^2)^{-1}$$

$$M^- \cos \varphi^- \rightarrow -\Omega^2 \omega_0^2$$

$$M^+ \sin \varphi^+ \rightarrow -\frac{1}{4} \alpha (\Omega^2 + \alpha^2)^{-1}$$

$$M^- \sin \varphi^- \rightarrow 2\alpha \omega_0^{-2}$$

and the coefficient  $M^- \cos \varphi^-$  becomes in (58) greater in modulus than the coefficient  $M^+ \cos \varphi^+$ .

It is observed that since in this case the quantities (59) are independent of  $k$ , the error associated with the term  $E_{ci}^{(2)}$  in eqns (45)–(49) proves equal to zero [this situation will occur every time that the adjustable parameters of filter  $\phi_2$  appear only in the numerator of the transfer function  $W_2(p)$ ].

Considering eqns (40), (41) and (28), it is noted that to increase the capability of the self-adjusting circuit for operating in non-quasi-stationary conditions one may use in the phase discriminators  $\Phi D_i$  the reference signals

$$m_{ic} \cos \Omega_i t + m_{is} \sin \Omega_i t \quad (60)$$

which are phase shifted with respect to the search modulation signal

$$\mu \Delta_i \cos \Omega_i t \quad (61)$$

In this case the processes of self-adjustment will proceed in accordance with the equations

$$\dot{X}_i = W_{\varphi i}(D) [E_t(\varepsilon^2(t) m_{ic} \cos \Omega_i t) + E_t(\varepsilon^2(t) m_{is} \sin \Omega_i t)] \quad (62)$$

where  $E_t(\varepsilon^2(t) m_{ic} \cos \Omega_i t)$  is determined by formulae (46)–(49), while

$$E_t(\varepsilon^2(t) m_{is} \sin \Omega_i t) = \frac{1}{4} \mu \Delta_i m_{is} (E_{si}^{(1)} + E_{si}^{(2)} + E_{si}^{(3)}) \quad (63)$$

$$E_{si}^{(1)} = \frac{\partial}{\partial X_i} \sum_{k=0}^M \theta_k^2 |W(j\omega_k)|^2 (M_{ik}^- \sin \varphi_{ik}^- - M_{ik}^+ \sin \varphi_{ik}^+) \quad (64)$$

$$E_{si}^{(2)} = - \sum_{k=0}^M \theta_k^2 |W(j\omega_k)|^2 \frac{\partial}{\partial X_i} (M_{ik}^- \sin \varphi_{ik}^- - M_{ik}^+ \sin \varphi_{ik}^+) \quad (65)$$

$$E_{si}^{(3)} = - \sum_{k=0}^M \theta_k^2 |W(j\omega_k)|^2 \frac{\partial \varphi(\omega_k)}{\partial X_i} (M_{ik}^+ \cos \varphi_{ik}^+ - M_{ik}^- \cos \varphi_{ik}^-) \quad (66)$$

Here the necessary condition (58) for normal operation of the self-adjusting circuits is replaced by the condition

$$M_{ik}^+ (m_{ic} \cos \varphi_{ik}^+ - m_{is} \sin \varphi_{ik}^+) + M_{ik}^- (m_{ic} \cos \varphi_{ik}^- + m_{is} \sin \varphi_{ik}^-) > 0 \quad (67)$$

which may prove much more favourable given a suitable choice of the phase of the voltage (60) (i.e. of the quantities  $m_{ic}$  and  $m_{is}$ ); the actual result of the undistorted forced process of self-adjustment comes out in this case to be the conditional minimization of the quantities

$$\sum_{k=0}^M \theta_k^2 |W(j\omega_k)|^2 [M_{ik}^+ (m_{ic} \cos \varphi_{ik}^+ - m_{is} \sin \varphi_{ik}^+) + M_{ik}^- (m_{ic} \cos \varphi_{ik}^- + m_{is} \sin \varphi_{ik}^-)] \quad (68)$$

In choosing the phase of the reference voltage (60) one can aim not only at increasing the coefficient (67) but also at the same time decreasing the quantity

$$|M_{ik}^+ (m_{ic} \sin \varphi_{ik}^+ + m_{is} \cos \varphi_{ik}^+) + M_{ik}^- (m_{ic} \sin \varphi_{ik}^- + m_{is} \cos \varphi_{ik}^-)| \quad (69)$$

i.e. (see (62), (66) and (49)) the error associated with the term  $E_{ci}^{(3)}$ . In practice, as a rule, it proves tedious to achieve an accurately optimum phase-shift (e.g. in the sense of a minimum ratio between the quantities (69) and (67)) between the signals (60) and (61), since by virtue of (29) this shift depends not only on the drifting parameters of filter  $\phi_2$  (a similar situation arises<sup>11</sup> also in extremal control systems), but also on the form of the test signal  $\theta(t)$ . Nevertheless by using *a priori* information on the operating conditions of the system, or by carrying out a running analysis of the signal  $\theta(t)$  and the results of system operation, in a number of cases one can evidently achieve an improvement in the dynamic properties of the given self-adjusting system relatively simply by using reference voltages of the form in (60) that only approximate to the optimum. In order to increase the stability of automatic phase-shift optimization between voltages (60) and (61) one can correlate the search and test signals in frequency [phase relations between the signals  $\theta(t)$  and  $\mu \Delta_i \cos \Omega_i t$  have no effect on the quantities (47)–(49), (64)–(66)].

The self-adjusting system, the phase of which use discriminators reference voltages of the general type given in (60) will be denoted by System B.

§ 11. The equations of motion (45)–(49) and (62)–(66) were derived under the assumption that the frequencies of the search modulation and the harmonic components of the test signal all satisfy the conditions (44). If these conditions do not hold, then the voltages  $E_t(\varepsilon^2(t) m_{ic} \cos \Omega_i t)$  and  $E_t(\varepsilon^2(t) m_{is} \sin \Omega_i t)$ , together with the signals (47)–(49) and (64)–(66), will also contain other components, which generally speaking will introduce certain additional distortions into the self-adjustment process. Equations (40) and (41) enable one effectively to calculate all these parasitic components of the control signal in the self-adjusting network.

For example, let only one of the conditions (44) be disturbed: let the frequency of the  $p$ th harmonic of the test signal coincide with the search frequency in the  $i$ th self-adjusting channel, i.e.  $\omega_p = \Omega_i$ . According to (41), in this case the signal  $E_t(\varepsilon^2(t) m_{ic} \cos \Omega_i t)$  will contain an additional term  $E_{ci}^{(4)}$ , generated by the presence in  $e_{kl}$  of harmonics with frequencies  $\omega_k + \omega_l - \Omega_q$  (for  $k = 0, l = p, q = i$  and  $k = p, l = 0, q = i$ ) and  $\omega_k + \omega_l - (\Omega_s + \Omega_q)$  (for  $k = l = p, s = q = i$ ):

$$\begin{aligned}
E_{ci}^{(4)} &= \frac{1}{2} m_{ic} E_t [\theta_0 \theta_p | W(0) W(j\omega_p) | (e_{po} + e_{op}) \\
&+ \frac{1}{2} \theta_p^2 | W(j\omega_p) |^2 e_{pp}] \cos \Omega_i t \\
&= m_{ic} \theta_0 \theta_p | W(0) W(j\omega_p) | \cos \vartheta_p \\
&+ \frac{1}{4} \mu \Delta_i m_{ic} \theta_p^2 | W(j\omega_p) |^2 x [a_i(\omega_p) \cos 2 \vartheta_p \\
&- b_i(\omega_p) \sin 2 \vartheta_p] \quad (70)
\end{aligned}$$

where  $\vartheta_p = \vartheta(\omega_p)$ ,  $a_i(\omega_p)$  and  $b_i(\omega_p)$  are defined respectively by eqns (36), (28) and (29) with  $\omega = \omega_p$ .

For the system considered in *Example 3*, the first term in expression (10) is zero (since  $\theta_0 = 0$ ), while the second may be calculated given the frequency characteristic of  $W_1(j\omega)$ . Even in this actual example it is, on the whole, difficult to judge what effect the use of a reference voltage of (60) type will have on the additional error in question. One can evidently achieve a stable reduction in this error or even its conversion into a useful signal, provided one correlates the search and test signals not only in frequency but also in phase, so as to limit unforeseen variations in the angle  $\vartheta_p$ .

#### IV. Calculation of Self-adjustment Operating Modes where the Test Action is a Stationary Random Process

§ 12. It is assumed for simplicity that the filters  $W_{\phi i}(D)$  in System A consist of elements which carry out the ideal averaging of the quantity  $m_{ic} \varepsilon^2(t) \cos \Omega_i t$  in time over the interval  $(t - T_0, t)$ :

$$W_{\phi i}(D) [\varepsilon^2(t) m_{ic} \cos \Omega_i t] = \frac{k_i}{T_0} \int_{t-T_0}^t \varepsilon^2(\tau) \cos \Omega_i \tau d\tau \quad (71)$$

and that the test signal  $\theta(t)$  is a time-function whose 'shortened' spectrum (10) actually only slightly depends on the instant of observation  $t$  and is located in the region of quite high frequencies:

$$\theta_T(j\omega, t) \approx \theta_T(j\omega), \quad \theta_T(j\omega) = 0 \text{ for } \omega < \omega^* \quad (72)$$

$$\omega^* - 2\Omega_N \geq T_0^{-1}, \quad \Omega_i - \Omega_{i-1} \gg T_0^{-1} \quad (73)$$

$$(\Omega_i > \Omega_{i-1}, \quad i = 1, \dots, N)$$

Every actual filter  $W(j\omega) = W_1(j\omega) - W_2(j\omega)$  has a finite cut-off frequency  $\omega_\phi$  (it is further considered that  $\omega^* < \omega_\phi$ ), so that in accordance with (19), (32), (34) and (71)–(73) the equations for the process of self-adjustment of the  $q$ th parameter may be put into the form

$$\dot{X}_q \approx \pi^{-2} \int_{\omega^*}^{\omega_\phi} d\omega \int_{\omega}^{\omega_\phi} dv T_0^{-1} \int_{t-T_0}^t G_q(\omega, v, \tau) d\tau \quad (74)$$

$$\begin{aligned}
G_q(\omega, v, \tau) &= D(\omega, v) [\cos \{(\omega - v)\tau + \vartheta_\omega - \vartheta_v\} \\
&+ \cos \{(\omega + v)\tau + \vartheta_\omega + \vartheta_v\}] \cos \Omega_q \tau
\end{aligned}$$

where  $D(\omega, v)$ ,  $\vartheta_\omega$  and  $\vartheta_v$  are defined by eqns (35), (36), (28) and (29). The quantity  $G_q(\omega, v, t)$  is a sum of harmonic components with frequencies  $\Omega$  equal to

$$\omega \pm v \pm \Omega_q, \quad \omega \pm v \pm (\Omega_s \pm \Omega_q), \quad (s = 1, \dots, N) \quad (75)$$

while the integral

$$\int_{t-T_0}^t G_q(\omega, v, \tau) d\tau$$

is a weighted sum of integrals of the type

$$\int_{t-T_0}^t \cos(\Omega\tau + \vartheta) d\tau \quad (76)$$

where  $\Omega$  are the frequencies in (75) and  $\vartheta$  are angles of the form  $\vartheta_\omega \pm \vartheta_v$  and  $\vartheta_\omega \pm \vartheta_v + \pi/2$ . On rewriting the integral (76) in the form

$$\int_{-\frac{1}{2}T_0}^{\frac{1}{2}T_0} \cos \left[ \Omega \left( \xi + t - \frac{1}{2}T_0 \right) + \vartheta \right] d\xi$$

it is observed that in accordance with a known<sup>8</sup> integral representation

$$\frac{1}{2} \pi^{-1} \int_{-\infty}^{\infty} \cos(\Omega\tau + \vartheta) d\tau = \cos \vartheta \cdot \delta(\Omega) \quad (77)$$

of the  $\delta$  function and for large enough averaging time intervals  $T_0$  of the filter  $W_{\phi q}(D)$ , the approximate equation

$$\begin{aligned}
\int_{t-T_0}^t \cos(\Omega\tau + \vartheta) d\tau &\approx \cos \left[ \Omega \left( t - \frac{1}{2}T_0 \right) + \vartheta \right] \delta(\Omega) \\
&= \cos \vartheta \cdot \delta(\Omega) \quad (78)
\end{aligned}$$

is true; using this, eqn (74) can be readily got into the form

$$\begin{aligned}
\dot{X}_q &\approx T \cdot T_0^{-1} \cdot k_q \cdot \pi^{-1} \int_{\omega^*}^{\omega_\phi} G_q(\omega) \\
&+ \frac{\mu}{2} \sum_{i=1}^N \Delta_i [g_i(\omega, v_{iq}^+) + g_i(\omega, v_{iq}^-) + g_q(\omega, v_{qq}^-)] d\omega \quad (79)
\end{aligned}$$

where

$$G_q(\omega) = |\theta_T(j\omega) \theta_T\{j(\omega + \Omega_q)\} W(j\omega) W\{j(\omega + \Omega_q)\}| \cdot T^{-1} \cos(\vartheta_\omega - \vartheta_{\omega + \Omega_q}) \quad (80)$$

$$\begin{aligned}
g_i(\omega, v) &= \frac{1}{2} T^{-1} |\theta_T(j\omega) \theta_T(jv) W(j\omega) W(jv)| \\
&\cdot [V_{ic}(\omega, v) \cos(\vartheta_\omega - \vartheta_v) - V_{is}(\omega, v) \sin(\vartheta_\omega - \vartheta_v)] \quad (81)
\end{aligned}$$

$$V_{ic}(\omega, v) = a_i(\omega) |W(j\omega)|^{-1} + a_i(v) |W(jv)|^{-1} \quad (82)$$

$$V_{is}(\omega, v) = b_i(\omega) |W(j\omega)|^{-1} + b_i(v) |W(jv)|^{-1}$$

$$v_{iq}^+ = \omega + \Omega_i + \Omega_q, \quad v_{iq}^- = \omega + |\Omega_i - \Omega_q| \quad (83)$$

[the quantities  $a_i(\omega)$ ,  $b_i(\omega)$  and  $\vartheta_\omega$  being defined by eqns (28), (29) and (36) and  $T$  the memory of filter  $W(j\omega)$ —see § 3].

Considering the function (5) as a typical realization of a stationary random process  $\{\theta(t)\}$  and performing averaging according to achievements, one can go from eqns (79)–(83) to equations in the mean (as taken together) values  $\bar{X}_q$  of the adjustable parameters. If here the interval  $T$  is taken large enough, then in the right-hand sides of these equations one may replace the quantities  $T^{-1} |\theta_T(j\omega) \theta_T(jv)|$  by characteristics like the mutual spectral power densities<sup>12</sup> of the process  $\{\theta(t)\}$  and certain random processes obtained from  $\{\theta(t)\}$  by simple transformations that do not infringe the stationary condition.

This paper does not deal with the more detailed analysis of the general case, but gives the results of the calculation for the quasi-stationary mode of self-adjustment, i.e. the mode in which

$$\omega^* > 2\Omega_N \quad (\Omega_i > \Omega_{i-1}, i = 1, \dots, N) \quad (84)$$

with a test signal of white-noise type:

$$T^{-1} |\theta_T(j\omega)|^2 \approx \lim_{T \rightarrow \infty} T^{-1} |\theta_T(j\omega)|^2 = \begin{cases} 0 & \text{for } \omega < \omega^* \\ G_0 & \text{for } \omega > \omega^* \end{cases} \quad (85)$$

Since eqns (30) and (31) are satisfied in quasi-stationary modes, and furthermore  $\vartheta_\omega \approx \vartheta_{\omega+2\Omega_M}$  ( $\omega > \omega^*$ ), one may neglect the terms  $V_{is}(\omega, \nu) \sin(\vartheta_\omega - \vartheta_\nu)$  in (81), and so putting  $\theta_T(j\omega) \approx \theta_T\{j(\omega + 2\Omega_M)\}$  and  $W(j\omega) \approx W\{j(\omega + 2\Omega_M)\}$ , the following equations for the self-adjustment process are arrived at:

$$\begin{aligned} \dot{\bar{X}}_q &\approx k_q^0 \left[ \int_{\omega^*}^{\omega_\phi} |W(j\omega)|^2 d\omega + \mu \Delta_q \frac{\partial}{\partial X_q} \int_{\omega^*}^{\omega_\phi} |W(j\omega)|^2 d\omega \right. \\ &\quad \left. + \frac{1}{2} \mu \sum_{\substack{i=1 \\ i \neq q}}^N \Delta_i \frac{\partial}{\partial X_q} \int_{\omega^*}^{\omega_\phi} |W(j\omega)|^2 d\omega \right] \quad (86) \\ k_q^0 &= T \cdot T_0^{-1} \cdot k_q^* \cdot \frac{1}{\pi} \cdot m_{qc} \cdot G_0 \end{aligned}$$

The following conclusions are evident from (86):

(1) In the mode of operation (84), (85) studied, minimization of the quantity

$$\int_{\omega^*}^{\omega_\phi} |W(j\omega)|^2 d\omega \quad (87)$$

may be naturally considered the ideal result of the self-adjustment process.

(2) The control signal for the  $q$ th self-adjusting network contains derivatives of the quantity (87) being minimized, not only w.r.t.  $X_q$  but also w.r.t. all the other adjustable parameters  $X_i$ , so that one has not got a pure gradient system of extremal control.

(3) The equilibrium condition  $\dot{\bar{X}}_q = 0$  ( $q = 1, \dots, N$ ) for the system (86) is characterized for  $\Delta_i = \Delta$  ( $i = 1, \dots, N$ ) by the relations

$$\mu \Delta (N+1) \frac{\partial}{\partial X_i} \int_{\omega^*}^{\omega_\phi} |W(j\omega)|^2 d\omega = - \int_{\omega^*}^{\omega_\phi} |W(j\omega)|^2 d\omega \quad (88) \quad (i = 1, \dots, N)$$

from which it can be seen that the strongest extremal nature of the dependence of quantity (87) on the parameters  $X_i$ , and the

less attainable the minimum of this quantity, the closer will this condition be to the ideal result of self-adjustment.

(4) If quite large differences arise rapidly between the frequency characteristics  $W_1(j\omega)$  and  $W_2(j\omega)$ , the non-negative term (87) on the right-hand side of eqn (86) will increase so much that the operation of the self-adjusting network will be reduced merely to increasing the parameter  $\bar{X}_q$  ( $\bar{X}_q > 0$ ), and this may lead to the system's losing its required extremal condition.

Finally it is noted that the equations given by Krasovski<sup>2</sup> for quasi-stationary self-adjustment with a white-noise test signal contain only terms analogous to the second term in the right-hand side of equation (86).

The author expresses his gratitude to Ye. A. Barbashin and I. N. Pechorina for their discussion of this paper.

## References

- 1 KRASOVSKI, A. A. Self-adjusting automatic control systems. *Automatic Control and Computer Engineering*. 1961. No. 4. Mashgiz
- 2 KRASOVSKI, A. A. The dynamics of continuous automatic control systems with extremal self-adjustment of the correcting devices. *Automatic and Remote Control*. 1960. London; Butterworths
- 3 KAZAKOV, I. YE. The dynamics of self-adjusting systems with extremal continuous adjustment of the correcting networks in the presence of random perturbations. *Automat. Telemekh.* 21, No. 11 (1960)
- 4 VARYGIN, V. N. Some problems in the design of systems with extremally self-adjusting correcting devices. *Automat. Telemekh.* 22, No. 1 (1961)
- 5 TAYLOR, W. K. An experimental control system with continuous automatic optimization. *Automatic and Remote Control*. 1960. London; Butterworths
- 6 MARGOLIS, M., and LEONDES, K. T. On the theory of self-adjusting control systems, the learning model method. *Automatic and Remote Control*. 1960. London; Butterworths
- 7 ITSKHOKI, YA. S. *Non-Linear Radio Engineering*. 1955. Sovetskoye Radio
- 8 KHÄRKEVICH, A. A. *Spectra and Analysis*. 1953. Gostekhizdat
- 9 MALKIN, I. G. *Some Problems in the Theory of Non-Linear Oscillation*. 1956. Gostekhizdat
- 10 POPOV, YE. P. *The Dynamics of Automatic Control Systems*. 1954. Gostekhizdat
- 11 CH'EN HSÜEH-SEN. *Technical Cybernetics*. 1956. Izd. Inostr. Lit.
- 12 LANING, G. H., and BETTIN, R. G. *Random Processes in Automatic Control Problems* (Russian transl.). 1958. Izd. Inostr. Lit.

# Some Considerations on Optimized Integrated Control

R. MARCACCI

## Summary

The authors have tried to find an analytical solution to the problem of linear integrated continuous control, of the  $n$ th order, in an open field, by optimizing its operation efficiency with the dynamic programming method and according to the point of view of Letov on the problem in question.

## Sommaire

Le rapport se propose de donner une solution analytique au problème d'un système de réglage linéaire continu d'ordre  $n$  en domaine ouvert en optimisant son fonctionnement par la méthode de la programmation dynamique, selon la méthode développée par Letov.

## Zusammenfassung

In dieser Arbeit wurde der Versuch unternommen, die analytische Lösung des Problems der gesamten kontinuierlich arbeitenden Regelung eines linearen vermaschten Systems  $n$ -ter Ordnung in einem offenen Bereich zu finden. Das Optimum der Regelgüte wird mit den Methoden des dynamischen Programmierens und nach einem von Letov vorgeschlagenen Verfahren gesucht.

## General

This paper is written with the object of optimizing a continuous control process, having a well-known linear differential transfer function, depending on several control variables defined in an open field according to the dynamic programming method applied by Bellman<sup>1</sup> and Letov<sup>2</sup>.

Suppose that a general differential transfer function be available within a process of the following type:

$$\dot{\eta}_k = G_k(\eta_1, \dots, \eta_n; \xi_1, \dots, \xi_m) \quad (1)$$

where  $k = 1, 2, \dots, n$  and  $G_k$  are continuous limited functions in  $\eta_1, \dots, \eta_n$  and  $\xi_1, \dots, \xi_m$ , allowing first partial derivatives with respect to  $\xi_1, \dots, \xi_m$ , and being defined in the open field  $N(\eta_1, \dots, \eta_n; \xi_1, \dots, \xi_m)$ .

$\eta_1, \dots, \eta_n$  are the process variables, considered as variations from the optimum prefixed values, and  $\xi_1, \dots, \xi_m$  are the control variables acting on the process variables  $\eta_1, \dots, \eta_n$ . For this reason they must more exactly be considered as unknown functions depending on  $\eta_1, \dots, \eta_n$ :

$$\xi_1 = \xi_1(\eta_1, \dots, \eta_n); \dots \xi_m = \xi_m(\eta_1, \dots, \eta_n)$$

The end of the assumed problem shall thus be to find out the

form of these functions  $\xi_1, \dots, \xi_m$  within a prefixed  $C$  class, by using the criterion of optimizing the operation efficiency of the control system.

In general, it might be assumed that all the process variables, on the whole, be controlled by an equal number of control variables, that is to say that  $m = n$ .

For the above reasons, it is clear that the control laws which are to be found,

$$\xi_1 = \xi_1(\eta_1, \dots, \eta_n); \dots \xi_m = \xi_m(\eta_1, \dots, \eta_n)$$

shall be those for which the optimal integrated process control is carried out on established values (prefixed setpoints) with a rigid feedback line. The resulting control action shall be that for which the system may reassume the stable equilibrium position as quickly as possible, whenever one or more simultaneous troubles shift it from such a position.

## Dynamic Programming

For readers' facility, the dynamic programming method is anticipated in order to obtain Bellman's functional equations. The control action of  $\xi_r$  ( $r = 1, 2, \dots, m$ ) must be such as to minimize the functional

$$I(\xi_1, \dots, \xi_m) = \int_0^\infty F(\eta_1, \dots, \eta_n; \xi_1, \dots, \xi_m) dt \quad (2)$$

defined in function class  $C(\xi_1, \dots, \xi_m)$  and in class  $C_1$  of variables  $\eta_1, \dots, \eta_n$ , where  $F$  is a function having a previously determined form. The minimization of (2) is equivalent to optimizing the operation efficiency of the control system.

Variable values in the limits shall be

$$\begin{array}{lll} \text{for } t_0 = 0 & \eta_s(t_0) = \eta_{s0} & \xi_r(t_0) = \xi_{r0} \\ \text{for } t_f = \infty & \eta_s(t_f) = 0 & \xi_r(t_f) = 0 \end{array}$$

where  $r = 1, 2, \dots, m$  and  $s = 1, 2, \dots, n$ .

Assume that  $\psi(\eta_{10}, \eta_{20}, \dots, \eta_{n0})$  be the functional  $I(\xi_1, \dots, \xi_m)$  minimized, that is to say

$$\psi(\eta_{10}, \dots, \eta_{n0}) = \min_{\xi_1, \dots, \xi_m} \int_0^\infty F(\eta_1, \dots, \eta_n; \xi_1, \dots, \xi_m) dt \quad (3)$$

Then it shall also be

$$\boxed{\text{Eqn (4)}}^*$$

where  $S$  is a small positive number.

\* Eqn (4):

$$\psi(\eta_{10}, \dots, \eta_{n0}) = \min_{\xi_1, \dots, \xi_m} \left[ \int_0^S F(\eta_1, \dots, \eta_n; \xi_1, \dots, \xi_m) dt + \int_S^\infty F(\eta_1, \dots, \eta_n; \xi_1, \dots, \xi_m) dt \right] \quad (4)$$

According to Bellman's optimality principle, the formula (4) changes into

$$\text{Eqn (5)}^*$$

and, if  $S$  is sufficiently small and if it is admitted that  $\psi$  be a continuous function, partially derivable within the range  $0 < S$ , one obtains the following formula:

$$\text{Eqn (6)}^\dagger$$

where  $0(S)$  is a residual, about which it may be supposed that the limit  $[0(S)/S] = 0$  for  $S \rightarrow 0$ , and  $\theta_K$  satisfies the formula  $0 \leq \theta_K < S$ . It results, from (6), that the following formula, obtained in the limit  $S \rightarrow 0$ , must be equally null:

$$0 = \min_{\xi_1, \dots, \xi_m} \left\{ F(\eta_{10}, \dots, \eta_{n0}; \xi_{10}, \dots, \xi_{m0}) + \sum_{k=1}^n \frac{\partial \psi}{\partial \eta_k} G_k \right\} \quad (7)$$

The formula (7) must be valid, whichever the initial time instant, thus:

$$\text{Eqn (8)}^\ddagger$$

which is fractionated into the following Bellman's system:

$$\left. \begin{aligned} \frac{\partial F}{\partial \xi_1} + \sum_{k=1}^n \frac{\partial \psi}{\partial \eta_k} \cdot \frac{\partial G_k}{\partial \xi_1} &= 0 \\ \dots \dots \dots \\ \frac{\partial F}{\partial \xi_m} + \sum_{k=1}^n \frac{\partial \psi}{\partial \eta_k} \cdot \frac{\partial G_k}{\partial \xi_m} &= 0 \end{aligned} \right\} \quad (9)$$

formed by  $(m+1)$  differential equations in  $(m+1)$  unknown variables  $\xi_1, \dots, \xi_m; \psi$ .

For the existence and unicity theorems of the solution  $\psi(\eta_1, \dots, \eta_n)$  of the system, see Bellman<sup>1</sup>. The general solution proceeding is to eliminate the unknown functions  $\xi_1, \dots, \xi_m$  in order to obtain a single differential equation which allows us to specify the solution  $\psi(\eta_1, \dots, \eta_n)$ . Then, by means of the last  $m$  equations, the  $m$  functions  $\xi_1, \dots, \xi_m$  are specified algebraically. By associating the resulting functions  $\xi_1, \xi_2, \dots, \xi_m$  with (1), it is possible to find out the time functions:

$$\eta_1 = \eta_1(t), \dots, \eta_n = \eta_n(t) \quad (10)$$

that is the laws of process variables under the optimized control action.

#### Study of Linear Integrated Control of $n$ order in an Open Field According to Letov

A transfer function within the process of the following linear type is assumed

$$\dot{\eta}_k = \sum_{\alpha=1}^n b_{k\alpha} \eta_\alpha + \sum_{\beta=1}^m m_{k\beta} \xi_\beta \quad (k=1, 2, \dots, n) \quad (1a)$$

with the initial statements:

$$\begin{aligned} \text{for } t_0=0 \quad \eta_k(0) &= \eta_{k0} & \xi_k(0) &= \xi_{k0} \\ \text{for } t_f=\infty \quad \eta_k(\infty) &= 0 & \xi_k(\infty) &= 0 \end{aligned}$$

and the function  $F$ , appearing in (2), of the following square positive definite type:

$$F = \sum_{s=1}^n a_s \eta_s^2 + \sum_{r=1}^m c_r \xi_r^2 \quad (11)$$

Then, Bellman's eqns (9) are specified as follows:

$$\left. \begin{aligned} \sum_{s=1}^n a_s \eta_s^2 + \sum_{r=1}^m c_r \xi_r^2 + \sum_{k=1}^n \frac{\partial \psi}{\partial \eta_k} \left[ \sum_{\alpha=1}^n b_{k\alpha} \eta_\alpha + \sum_{\beta=1}^m m_{k\beta} \xi_\beta \right] &= 0 \\ 2 c_r \xi_r + \sum_{k=1}^n \frac{\partial \psi}{\partial \eta_k} m_{kr} &= 0 \quad (r=1, 2, \dots, m) \end{aligned} \right\} \quad (9a)$$

From the second equation of (9a) it results that  $\xi_r$  are bound by relations:

$$\xi_r = -\frac{1}{2c_r} \sum_{k=1}^n m_{kr} \frac{\partial \psi}{\partial \eta_k}, \quad (r=1, 2, \dots, m) \quad (9b)$$

where the existence and unicity of solution for  $\xi_r$  clearly appear in an open field, as soon as the existence and unicity of solution of functional  $\psi$  are assured.

By replacing (9b) in the first of (9a), the differential equation, providing that the function  $\psi$  is known, is obtained. The equation is the following one:

$$\begin{aligned} \sum_{s=1}^n a_s \eta_s^2 + \sum_{r=1}^m c_r \cdot \frac{1}{4c_r^2} \left( \sum_{k=1}^n \frac{\partial \psi}{\partial \eta_k} m_{kr} \right)^2 \\ + \sum_{k=1}^n \frac{\partial \psi}{\partial \eta_k} \left[ \sum_{\alpha=1}^n b_{k\alpha} \eta_\alpha - \sum_{\beta=1}^m m_{k\beta} \frac{1}{2c_\beta} \sum_{v=1}^m \frac{\partial \psi}{\partial \eta_v} m_{v\beta} \right] &= 0 \end{aligned} \quad (12)$$

and simplifying:

$$\sum_{s=1}^n a_s \eta_s^2 + \sum_{k=1}^n \frac{\partial \psi}{\partial \eta_k} \sum_{\alpha=1}^n b_{k\alpha} \eta_\alpha = \frac{1}{4} \sum_{k=1}^n c_k \left( \sum_{r=1}^m \frac{\partial \psi}{\partial \eta_r} m_{rk} \right)^2 \quad (12a)$$

The solution shall be of the following type:

$$\psi = \sum_{k=1}^n \sum_{\alpha=1}^n A_{k\alpha} \eta_k \eta_\alpha \quad (13)$$

and its coefficients  $A_{k\alpha}$  may be obtained by well-known methods.

Calculating all the partial derivatives  $\partial \psi / \partial \eta_k$ , it is possible, by means of (9b), to obtain the functions  $\xi_r$  ( $r=1, 2, \dots, m$ ), which shall be of the following type:

$$\xi_r = \sum_{\alpha=1}^n p_{r\alpha} \eta_\alpha, \quad (r=1, 2, \dots, m) \quad (14)$$

These represent the law of optimized integrated proportional control, with fixed band. Control proportionality results, dependent on the particular choice of functional  $F$ .

$$^* \text{Eqn (5):} \quad \psi(\eta_{10}, \dots, \eta_{n0}) = \min_{\xi_1, \dots, \xi_m} \left[ \int_0^S F(\eta_1, \dots, \eta_n; \xi_1, \dots, \xi_m) dt + \psi(\eta_1(S), \dots, \eta_n(S)) \right] \quad (5)$$

$$^\dagger \text{Eqn (6):} \quad \psi(\eta_{10}, \dots, \eta_{n0}) = \min_{\xi_1, \dots, \xi_m} \left[ F(\eta_{10}, \dots, \eta_{n0}; \xi_{10}, \dots, \xi_{m0}) S + \psi(\eta_{10}, \dots, \eta_{n0}) + \left\{ \sum_{K=1}^n \left( \frac{\partial \psi}{\partial \eta_K} G_K \right)_{0K} \right\} S + 0(S) \right] \quad (6)$$

$$^\ddagger \text{Eqn (8):} \quad 0 = \min_{\xi_1, \dots, \xi_m} \left\{ F(\eta_1, \dots, \eta_n; \xi_1, \dots, \xi_m) + \sum_{k=1}^n \frac{\partial \psi}{\partial \eta_k} G_k(\eta_1, \dots, \eta_n; \xi_1, \dots, \xi_m) \right\} \quad (8)$$

Associating (14) with (1a), the law of  $\eta_s = \eta_s(t)$ , ( $s = 1, \dots, n$ ) may also be determined. The same result might be reached with the help of Lagrange calculus of variations.

The control system obtained by (14) is also stable, once it is possible to assure the positivity of function  $\psi$  in

$$\sum_{s=1}^n \eta_s^2 = 0$$

This is obtained according to Letov's point of view<sup>2</sup>. Then this function  $\psi$  coincides with a Liapunov's function.

#### References

- <sup>1</sup> BELLMAN, R. *Dynamic Programming*. 1957. Princeton; Princeton University Press
- <sup>2</sup> LETOV, A. M. *Analytical Design of Controllers*. Sec IV, Russian Academy of Sciences; *Automat. Telemekh.* 22, No. 4 (1961)

#### DISCUSSION

A. LETOV, *Institute of Automation and Telemekhanics, Moscow, U.S.S.R.*

The problem dealt with in the paper is quite interesting. However, when it is considered in a closed space it gives rise to another purely mathematical problem, the solution of which is a necessary condition of existence of the optimal solution found for  $\xi_r$  ( $r = 1, \dots, m$ ).

The function generating the solution  $\xi_r$  should be at least continuous on the boundaries of the closed space.

In the simplest case  $r = 1$ , when the closed space of the problem is defined by the inequality  $|\xi_1| \leq 1$ , the problem of the continuity of the generating function is reduced to the solution of the Cauchy problem for a specially constructed partial equation following from Bellman's method.

This problem has a positive solution in the given simple case. It would be extremely interesting to know the solution of the given problem in the case  $r > 1$ .

S. KAHNE, *U.S.A.F., Cambridge Research Laboratories, Bedford, Mass., U.S.A.*

To solve a problem in optimal control theory Dr. Marcacci has used Bellman's notion of dynamic programming. For linear systems with a quadratic performance functional, his procedure leads, in a straightforward manner, to an optimal control law which is a linear combination of the state variables. I interpret the author's 'optimized integrated proportional control, with fixed band' to be a suggestion for a feedback solution to the problem. It should be noted that if one or more states are not observable, the solution would have to be modified. I might suggest, therefore, that the class of problems for which his eqn (14) is a valid feedback solution is the class of completely observable linear systems with performance functional  $F$  of his eqn (11).

The case considered in the paper resulted in a non-linear partial differential equation for which a solution could be found by inspection. In the more general case where the system or the performance functional was of a different form it would seem that this solution guessing would be difficult indeed. I would appreciate the author's comments about the possibility of extending his method to these more general situations.

# Optimal Processes in Systems with Time Lag

N. N. KRASOVSKII

## Summary

Problems are considered of constituting the controller action in a system with time lags of pulses in the plant and of signals in the feedback channels. The problems are complicated by random load and disturbances. Statements of the problems are given, and criteria of optimality based on Liapunov's method of functions, worked out for systems with after-action, and brought up to date in accordance with the principles of *dynamic programming*. An explicit form of optimal control is derived for the problem of the analytic design of a regulator to minimize a quadratic functional; in this example optimal control is put together from a linear functional of the prehistory of the process and a term determined by the actual value of the load. Approximate methods of calculation of the optimal control signal are discussed, and generalizations of the problem are indicated. A class of systems with incomplete information is indicated for which the problem of the analytic design of a regulator can be broken down, thus: (1) optimal control is found for a similarly determined problem; (2) The best prognosis of the state of the plant and load is established according to the given feedback channels; (3) the prognosis of (2), introduced into the law of control (1), gives optimal control.

## Sommaire

Il s'agit des problèmes liés à la génération d'une action de commande dans un système dans lequel il existe des retards entre les impulsions données à l'installation et les signaux cheminant le long des circuits de réaction. L'existence d'une charge et de perturbations aléatoires complique les problèmes. La communication précise les énoncés des problèmes ainsi que les critères d'optimalité; ces derniers sont établis à l'aide de la méthode des fonctions de Liapunov, adaptée aux systèmes à action résiduelle et modernisée conformément aux principes de la *programmation dynamique*. On en déduit une forme explicite de commande optimale qui est adaptée au problème de l'étude analytique d'un régulateur capable de minimiser une fonctionnelle quadratique. La commande optimale résulte de la superposition d'une fonctionnelle linéaire de la préhistoire du processus et d'un terme dépendant de la valeur effective de la charge. On donne des méthodes permettant le calcul approché du signal de commande optimal et on indique la généralisation du problème. On montre, que pour une classe de systèmes à informations incomplètes, le problème de la synthèse analytique d'un régulateur peut être décomposé en deux: (1) on conçoit la commande optimale par analogie avec un problème similaire; (2) compte tenu des circuits donnés de réaction, on fait la meilleure prédiction quant à l'état de l'installation et quant à sa charge. Le fait d'introduire les valeurs ainsi prédites (2) dans la loi de commande (1) conduit à la commande optimale.

## Zusammenfassung

Der Aufsatz behandelt den Entwurf von Reglern für Systeme mit Zeitverzögerungen in der Strecke und mit verzögerter Rückführung. Die als regelloses Signal auftretende Last oder Störungen erhöhen die Schwierigkeiten. Es werden die Probleme formuliert, optimale Bedingungen aufgrund der Methode der Liapunov-Funktionen für Systeme mit „Nachwirkung“ ausgearbeitet und die Lösungsform in Übereinstimmung mit den Prinzipien der dynamischen Programmierung auf den neuesten Stand gebracht. Für das Problem des analytischen Entwurfs eines Reglers zur Minimierung eines quadratischen Funktionals wird eine explizite Darstellung der optimalen Regelung abgeleitet. Hierbei ergibt sich die optimale Regelung aus einem linearen Funktional der Vorgeschichte des Prozesses und einem Ausdruck, der

durch die wirklichen Werte der Last bestimmt ist. Näherungsmethoden zur Berechnung der optimalen Regelgröße werden besprochen und Verallgemeinerungen des Problems aufgezeigt. Die Arbeit betrachtet eine Gruppe von Systemen mit unvollständiger Information, für die sich das Problem des analytischen Entwurfs des Reglers in drei Teile zerlegen läßt: 1. Zunächst sucht man die optimale Regelung eines ähnlichen Problems. 2. Aufgrund der gegebenen Rückführpfade wird die beste Vorhersage des Zustandes der Strecke und der Last gewonnen. 3. Die Vorhersage von 2. in die Beziehungen der Regelung 1. eingesetzt, ergibt die optimale Regelung.

## Introduction

The problem of forming the optimal process input for a regulator in a system with time lag of action and signals is considered in this paper. The questions considered belong to the class of problems of optimal control. These problems were first stated and developed in the U.S.S.R by Feldbaum<sup>1</sup>. The mathematical theory of optimal processes was worked out by Pontryagin *et al.*<sup>2</sup>, on the basis of their Maximum Principle. Their studies have given rise to a great number of works: for instance, that by Rozonoer<sup>3</sup>, and also the *Theory of Dynamic Programming*<sup>4</sup>, developed by Bellman and his colleagues on the basis of the optimality principle and the functional equations which follow from it, which embraces a very wide class of problem. Reference can be made to the authors whose works, among others, have a direct connection with this paper<sup>5-18</sup>.

Reference can also be made to the works of those authors who, among others, have studied optimal control problems in after-action systems, and in more general systems with distributed parameters<sup>19-22</sup>.

The present work originates from the studies of Letov<sup>23, 24</sup>, and the statement of the problem adopted here is a generalization, for systems with after-action, of the statement of the problem given by Letov<sup>24</sup>. The problems for systems with delay of the feedback signals considered below are related to problems of dual control<sup>25</sup> or of the theory of adaptive processes<sup>26</sup>.

The solution proposed is based on the method of Liapunov functions and the theory of stability of motion<sup>27, 28</sup>, developed for equations with time lags<sup>29</sup>, and modernized in accordance with the principles of *Dynamic Programming*<sup>4</sup>. Statements of the problems are given in this paper, and criteria of optimality and the principles of solution are formulated. For systems which can be described by a few actual equations, the explicit analytical form of the optimal regulator is given. Approximate methods for calculating optimal control are described, and problems complicated by random circumstances considered.

## Time-lag of Signals in the Plant

Consider a controlled system (Figure 1) where  $z(t)$  is a controlled vector quantity at the output of the plant  $A$ , and  $\xi$ , a

scalar quantity, is the input of the regulator  $B$ , constituted on the basis of information on the actual error  $x = z - z^0$ , and possibly also on the actual values of the load  $\eta(t)$ . The special feature of the system is the time-lag of the signals in the plant  $A$  (Case I), or of signals in the feedback channels (1) and (2) (Case II), or of  $\xi$  in channel (3) (Case III). Each case will be examined separately. If Cases I—III are combined in one system, the statement of the problems and the solutions must be combined accordingly.

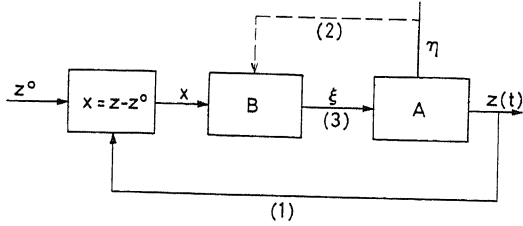


Figure 1

*Case I.* Assume that the disturbed motion of the system is described by the equation

$$\frac{dx}{dt} = f[t, x(t), x(t-h_1), \dots, x(t-h_k), \eta(t), \xi] \quad (1)$$

where  $x$  is an  $n$ -dimensional error vector,  $h_i$  is the time lag of signals in the plant ( $0 < h_i \leq h$ ,  $i = 1, \dots, k$ ),  $f$  is a known vector function of its own arguments, determined by the structure of the system, and  $\eta(t)$  is the load or disturbance. Besides this, a functional determining the quality of the process is given, and there may be a restriction on the magnitude of the control signal  $\xi$ .

The disturbed motion  $x(t)$  of system (1) with after-action, with  $t > t_0 \geq 0$  is determined, as is well known, by the history  $x(t_0 + \theta)$  ( $-h \leq \theta \leq 0$ ) of this motion. The initial function  $x(t_0 + \theta)$  ( $-h \leq \theta \leq 0$ ) will therefore be called the initial disturbances (with  $t = t_0$ ). It is also convenient to consider, as quantities describing the state of system (1) at instants  $t \geq t_0$ , and determining its future motion when  $\tau > t$ , sections of the trajectories  $x(t + \theta)$  ( $-h \leq \theta \leq 0$ ). It is therefore suitable to form the control signal  $\xi(t)$  at each instant  $t$  on the basis of information on the whole of the realized trajectory  $x(t + \theta)$  with  $-h \leq \theta \leq 0$ . In other words, analytic construction of the regulator<sup>24</sup> means finding  $\xi$  in the form of a some functional  $\xi(t) = \xi[t, x(t + \theta)]$ , determined on the curves  $x(t + \theta) = \{x_i(t + \theta), -h \leq \theta \leq 0, i = 1, \dots, n\}$ . In future it will be assumed that the argument  $\theta$  varies within the limits  $-h \leq \theta \leq 0$ . The continuous functions  $x(\theta)$  or  $x(t + \theta)$  of the argument  $\theta$  are assumed to be elements of a certain space  $X$  with a norm

$$\|x(\theta)\| = \max_{\theta} (x_1^2(\theta) + \dots + x_n^2(\theta))^{\frac{1}{2}}$$

Also used is the notation

$$\|x(0)\| = (x_1^2(0) + \dots + x_n^2(0))^{\frac{1}{2}},$$

$$\|x(t)\| = (x_1^2(t) + \dots + x_n^2(t))^{\frac{1}{2}}$$

Three problems are considered:

*Problem 1.* Find a control signal  $\xi = \xi^0[t, x(\theta)]$  such that the motion  $x = 0$  in a closed system (1), that is, with  $\xi(t) = \xi^0[t, x$

$(t + \theta)]$  is asymptotically stable<sup>29</sup> with respect to the disturbances  $x^0(t_0 + \theta)$  ( $t_0 \geq 0$ ) from a region

$$\|x^0(\theta)\| \leq G_0 \quad (2)$$

and such that for all  $t_0 \geq 0$  and  $x^0(t_0 + \theta)$  out of (2) there holds a minimum

$$J[t_0, x^0, \xi^0] = \min_{\xi} J[t_0, x^0, \xi] \quad (3)$$

Here

$$J[t_0, x^0, \xi] = \int_{t_0}^{\infty} \omega[t, x(t, t_0, x^0, \xi), \xi(t)] dt \quad (4)$$

where  $\omega$  is a given non-negative function,  $x(t, t_0, x^0, \xi)$  is the trajectory of (1) with initial conditions  $t_0$  and  $x^0(t_0 + \theta)$  and a selected law of control  $\xi(t) = \xi[t, x(t + \theta)]$ . The control signal  $\xi$  can be constrained by a supplementary restriction  $\xi \in \Xi$  (for instance,  $|\xi| \leq 1$ ).

*Problem 2.* Find a control signal  $\xi = \xi^0[t, x(\theta)]$  assuring a minimum of

$$J_T[t_0, x^0, \xi^0] = \min_{\xi \in \Xi} J_T[t_0, x^0, \xi] \quad (0 \leq t_0 \leq T) \quad (5)$$

where

$$J_T[t_0, x^0, \xi] = \int_{t_0}^T \omega[t, x(t, t_0, x^0, \xi), \xi(t)] dt + \psi[x(T, t_0, x^0, \xi)] \quad (6)$$

and  $T < \infty$  is a given instant of time, while  $\|x^0(t_0 + \theta)\| \leq G_0$ .

*Problem 3.* Find a control signal  $\xi = \xi^0[t, x(\theta)]$  assuring minimum of

$$J_{\infty}[t_0, x^0, \xi^0] = \min_{\xi \in \Xi} J_{\infty}[t_0, x^0, \xi] \quad (7)$$

where  $\|x^0(t_0 + \theta)\| \leq G_0$  and

$$J_{\infty}[t_0, x^0, \xi] = \lim_{T \rightarrow \infty} \frac{J_T}{T - t_0} \quad (8)$$

In Problems 2 and 3, as in 1, it is assumed that the initial conditions  $x^0$  and trajectories  $x(t, t_0, x^0, \xi^0)$  do not go beyond certain previously fixed regions.

The sufficient conditions of optimality of the control signal  $\xi^0$  will be formulated for Problems 1 and 2.

*Theorem 1.* Let it be possible to indicate functionals  $v[t, x(\theta)]$  and  $\xi^0[t, x(\theta)]$ , defined and satisfying in some region  $\|x(\theta)\| \leq G$  the following conditions:

(1) The functional  $v$  is positive definite with respect to  $\|x(0)\|$ .

(2) The functional  $v$  admits an upper limit with respect to  $\|x(\theta)\|$ .

(3) The following inequality is satisfied:

$$\inf [v[t, x(\theta)] \text{ when } \|x(0)\| = G,$$

$$\|x(\theta)\| = G] \geq \sup [v[t, x(\theta)] \text{ when } \|x(\theta)\| \leq G_0]$$

(4) Along trajectories of (1)<sup>29</sup> the derivative  $(dv/dt)_{\xi}$  of the functional  $v$  satisfies the condition

$$\left(\frac{dv}{dt}\right)_{\xi^0} + \omega[t, x(t), \xi^0] = \min_{\xi \in \Xi} \left[ \left(\frac{dv}{dt}\right)_{\xi} + \omega[t, x(t), \xi] \right] = 0 \quad (9)$$



in the region  $\|x(t + \theta)\| \leq G$ , and is negative definite with respect to  $\|x(t)\|$  in this region.

Then  $\xi^0[t, x(t + \theta)]$  is the optimal control signal for Problem 1, and the following equality is valid:

$$v[t_0, x^0(t_0 + \theta)] = J[t_0, x^0(t_0 + \theta), \xi^0] \quad (10)$$

*Note.* Properties (1) and (2) generalize in a natural way the corresponding properties of Liapunov's functions<sup>27</sup> that is (1) means that there exists a function  $w(r) > 0$  with  $r \neq 0$ , such that  $v[t, x(\theta)] \geq w(\|x(0)\|)$  with  $\|x(\theta)\| = \|x(0)\|$ , and (2) means that there exists a function  $W(r)$  satisfying the conditions  $W(0) = 0$ ,  $v[t, x(\theta)] \leq W(\|x(\theta)\|)$ . If in Problem 1 the region  $G_0$  encompasses any possible large initial disturbances  $x_0$  (the problem of optimal stabilization as a whole), the region  $G$  must coincide with the whole of the space  $X$ , and (1) is replaced by the condition

$$\lim v[t, x(\theta)] = \infty \text{ when } \|x(0)\| \rightarrow \infty, \|x(\theta)\| = \|x(0)\| \quad (11)$$

uniformly with respect to  $t$ .

The demonstration of Theorem 1 is made by reasoning typical for the theory of stability of motion<sup>29</sup>, but taking into account the principles of dynamic programming<sup>4</sup>.

The sufficient criterion of optimality for Problem 2 is formulated as follows:

*Theorem 2.* Let there exist for every  $\|x^0(t_0 + \theta)\| \leq G_0$  and  $t_0 \in [0, T]$  an admissible control signal  $\xi(t)$ , that is, a control signal for which the trajectory  $x(t, t_0, x^0, \xi)$  may be prolonged in some finite region  $G$  until the instant  $t = T$ , and therefore the integral (6) is finite. If one can find in the region  $G$  functionals  $v[t, x(\theta)]$  and  $\xi^0[t, x(\theta)]$  satisfying conditions (9), and

$$v[T, x(\theta)] = \psi[x(0)] \quad (12)$$

then  $\xi^0$  is the optimal control signal for Problem 2, and the following equality is valid:

$$v[t_0, x^0(t_0 + \theta)] = J_T[t_0, x^0(t_0 + \theta), \xi^0] \quad (13)$$

The solution of Problem 3 can be obtained by passage to the limit from the solution of the problem when  $T \rightarrow \infty$ .

*Note.* If the load  $\eta(t)$  is random or the system is subject to random disturbance, Problems 1 to 3 are modified as follows: integrals (4), (6) and (8) are replaced by their mathematical expectations (the conditional mathematical expectations for the appropriate initial conditions  $t_0, x^0, \eta^0$ ), and in Problem 1 the requirement of stability is replaced by the requirement of stochastic stability<sup>30</sup>. In this case seek the control signal  $\xi^0$  in the form of a functional  $\xi^0[t, x(t + \theta), \eta(t + \tau)]$ , where  $-h \leq \theta \leq 0$  and  $-h^* \leq \tau \leq 0$ , while  $h^* \geq 0$  is the value of the maximal after-action for the probability process  $\eta(t)$  (if  $\eta(t)$  is a Markov process, then  $h^* = 0$ ). The criteria of optimality given above preserve their form, with the modification that  $v$  must here also be a functional  $v[t, x(\theta), \eta(\tau)]$ , and the derivative  $(dv/dt)_\xi$  is replaced by its average value<sup>30</sup>  $(dM\{v\}/dt)_\xi$ .

Conditions (9) reduce to partial derivative equations of a special kind. The solution of these equations in the general case is cumbersome; it is possible, however, to indicate a number of cases when an explicit form can be found for the optimal control signal, or when a numerical procedure for its determination can be indicated.

The results of applying the proposed criteria to systems described by equations of actual form will be illustrated.

Let the transient process be described by the linear differential equations

$$\frac{dx_i}{dt} = \sum_{j=1}^n a_{ij}(t)x_j(t) + \sum_{j=1}^n c_{ij}(t)x_j(t-h) + b_i\xi + a_i\eta(t) \quad (14)$$

where  $a_{ij}, c_{ij}, a_i$  and  $b_i$  are known functions of time or constants. First assume that  $\eta(t) \equiv 0$ , and then consider Problem 1 for system (14), assuming that

$$J = \int_{t_0}^{\infty} \left[ \sum_{i=1}^n x_i^2(t) + \lambda \xi^2(t) \right] dt, \quad \lambda > 0 - \text{const} \quad (15)$$

any initial disturbances  $x^0(t_0 + \theta)$  are admissible.

Here the functional  $v$  from Theorem 1 must be chosen in the form

$$\begin{aligned} v[t, x(\theta)] = & \sum_{i,j=1}^n [d_{ij}(t)x_i(0)x_j(0) \\ & + 2x_i(0) \int_{-h}^0 \beta_{ij}(t, \theta)x_j(\theta) d\theta \\ & + \int_{-h}^0 \int_{-h}^0 \gamma_{ij}(t, \theta, \tau)x_i(\theta)x_j(\tau) d\theta d\tau] \end{aligned} \quad (16)$$

which generalizes in a natural way the Liapunov function widely used in stability theory, as a quadratic form. If for every initial condition  $x^0, t_0$  there exists an admissible control signal  $\xi(t)$ , that is, a control signal  $\xi(t)$  for which integral (15) converges uniformly with respect to  $t_0$ , then there exists a functional  $v$  (16) satisfying the conditions of Theorem 1. From this it is directly concluded that in this case there exists an optimal control signal  $\xi^0$  having the form

$$\begin{aligned} \xi^0[t, x(t + \theta)] \\ = \sum_{i=1}^n \left[ \mu_i(t)x_i(t) + \int_{-h}^0 v_i(t, \theta)x_i(t + \theta) d\theta \right] \end{aligned} \quad (17)$$

## Conclusion

The optimal regulator  $\xi^0$  in system (14) with condition of minimum (15) is seen to be the regulator  $B$ , which applies to the input of the controlled plant  $A$  at every instant  $t$  a quantity  $\xi^0$  (17), worked out on the basis of a measurement of the error  $x$  at the given instant of time  $t$  and at previous instants  $t - h \leq \tau \leq t$ , while the results of measurement of the previous errors  $x(\tau) = x(t + \theta)$  must be processed in the integrators  $\int v_i(t, \theta)x_i(t + \theta) d\theta$ . The control signal  $\xi^0$  depends linearly on  $x(t + \theta)$  ( $-h \leq \theta \leq 0$ ).

It is interesting to observe that for a system (14) with discrete delay  $h > 0$  the optimal control signal must be worked out by an element with continuous distribution of the after-action  $v_i(t, \theta)$  over the whole of the time-lag interval  $-h \leq \theta \leq 0$ .

Now let  $\eta(t) \equiv 0$  be a known function of time. Consider for system (14) the problem (2), where

$$J_T = \int_{t_0}^T \left[ \sum_{i=1}^n x_i^2(t) + \lambda \xi^2(t) \right] dt + \sum_{i,j=1}^n \psi_{ij}x_i(T)x_j(T) \quad (18)$$

Here any restricted control signal  $\xi(t)$  is admissible, and the following assertion is valid: a functional  $v$  satisfying the

conditions of Theorem 2 exists, and differs in form from the functional (16) by the term

$$v^* = \sum_{i=1}^n (\delta_i(t) x_i(0) + \int_{-h}^0 \varphi_i(t, \theta) x_i(\theta) d\theta) + \varrho(t) \quad (19)$$

From this assertion follows the conclusion that in this case an optimal control signal always exists, and differs in form from the control signal (17) by a term  $\xi^* = \kappa(t)$  which is a function only of time  $t$ .

*Note.* The conditions for Problem 1 solvability for systems (14) and (15) reduce to the possibility of constructing an admissible control signal  $\xi(t)$ . Here, as also in the case of systems without delay, the question is connected with the conditions of controllability of the system<sup>11, 31</sup>. System (14) (with  $\eta(t) \equiv 0$ ) will be called fully controlled in the interval  $[t_0, t_1]$  ( $t_1 > t_0 + h$ ) provided that for every initial condition  $x^0(t_0 + \theta)$  there exists a continuous (piece-wise-continuous) control signal  $\xi(t)$  such that  $x(t, t_0, x^0, \xi) \equiv 0$  when  $t_1 - h \leq t \leq t_1$ . The conditions of controllability, as in the case without delay<sup>31</sup>, can be investigated starting from the 'L problem'. If system (14) is fully controllable in every sufficiently long section of the  $t$  axis, then it is optimally stabilizable in the sense of Problem 1. It is also observed that such stabilization is certainly possible if system (14) is asymptotically stable with  $\xi = 0$ , or if the delay  $h > 0$  is sufficiently small (or if the  $c_{ij}$  are small), and for the system  $dx_i/dt = \sum a_{ij} x_j + b_i \xi$  the conditions of full controllability are fulfilled: the vectors  $\{b_i\}$ ,  $\{\|a_{ij}\| \{b_i\}\}$ , ...,  $\{\|a_{ij}\|^{n-1} \{b_i\}\}$  are linearly independent. The conditions of solvability of Problem 1 for (14) and (15) can also be ascertained in the process of solution, if the solution is sought by passage to the limit from the solution of Problem 2 for (14) and (18) (with  $\psi_{ij} = 0$ ,  $\eta = 0$  and with  $T \rightarrow \infty$ ), which is sometimes a convenient method in practice.

Now consider Problem 3 for system (14): accept that in (8)  $\omega = \sum_{i=1}^n x_i^2 + \lambda \xi^2$  and assume  $\eta(t)$  to be a random Markov function (for definiteness, of the pure discontinuous or diffusion type). Moreover, assume that system (14) is subject to some irregular disturbance of the white noise type, causing diffusion spread of  $x(t)$  in the time  $dt$  with a matrix of second moments  $\|M\{dx_i dx_j\} = \|\sigma_{ij}(t) dt\|$ .

The following result is obtained: if system (14) with  $\eta(t) \equiv 0$  is stabilized in the sense of Problem 1, then an optimal control signal  $\xi^0$  exists and has the form

$$\xi^0[t, x(t+\theta), \eta(t)] = \sum_{i=1}^n \left[ \mu_i(t) x_i(t) + \int_{-h}^0 v_i(t, \theta) x_i(t+\theta) d\theta \right] + \kappa(t, \eta(t)) \quad (20)$$

It is interesting to observe that the first term here tallies with (17), and the random term  $\kappa(t, \eta(t))$  determined by the actual values of  $\eta(t)$  is the same as it would be if, with  $\tau > t$ , the function  $\eta(\tau)$  were determined and tallied with the prediction of its mathematical expectation  $M\{\eta(\tau)/\eta(t)\}$  made according to the actual value of  $\eta(t)$ . The magnitude of the dispersion of  $\eta(t)$  and the quantities  $\sigma_{ij}(t)$  do not affect  $\xi^0$ , and manifest themselves only naturally in the quantity  $M\{J_\infty[t_0, x^0, \eta^0, \xi^0]\}$ .

As has been shown above, it is very laborious, in the general case, to construct the functional from Theorems 1 and 2. The

following methods may be indicated for its approximate determination (and consequently that of  $\xi^0$ ): the small parameter method; approximate solution of the functional equation (9); approximating  $v$  in the mean; replacing the equations with delays or the functional equation (9) by finite difference equations; replacing the equation with delays by a set of equations for the Fourier coefficients of a section of the trajectory  $x(t+\theta)$ ,  $-h \leq \theta \leq 0$ . These methods can be illustrated by numerical examples.

### Delay of Feedback Signals

Consider now the system of Figure 1 when there is no after-action in the plant  $A$ , but signals in channels 1—3 can be delayed.

*Case II.* Let the motion of the plant  $A$  be described by the vector differential equation

$$\frac{dx}{dt} = f[t, x(t), \eta(t), \xi] + \phi \quad (21)$$

where  $x, \eta, \xi, f$  have the same meaning as in the first part of the paper, and  $\phi$  is a disturbance of the white noise type, giving rise to diffusion spread of  $x(t)$  in the time  $dt$  with the matrix

$$\|M\{dx_i dx_j\} = \|\sigma_{ij}(t)\| dt \quad (22)$$

The problem is to minimize the quantities

$$J_T = M \left\{ \int_{t_0}^T \omega[t, x(t), \xi(t)] dt + \psi[x(T)] \right\} dt \quad (23)$$

and

$$J_\infty = \lim_{T \rightarrow \infty} \frac{J_T}{T - t_0} \text{ with } T \rightarrow \infty \quad (24)$$

The peculiarity of the case in question is that information concerning the actual values of the error  $x(t)$  and load  $\eta(t)$  are supplied by way of channels 1 and 2 with delays of  $h_1 > 0$  and  $h_2 > 0$  (or either  $h_1 > 0$  or  $h_2 > 0$ ) respectively ( $h_1 \leq h, h_2 \leq h$ ). In other words, assume that in the regulator  $B$  at the instant  $t$  in the closed interval  $[0, T]$  the values of the actual quantities  $x(t-h_1)$  and  $\eta(t-h_2)$ , where  $\eta(t)$  is a random Markov function, are known. Also assume that the regulator  $B$  is capable of remembering up to the instant  $t$  the signal  $\xi(t+\theta)$  worked out by it with  $-h \leq \theta < 0$ . Denote the set of magnitudes  $x(t-h_1)$ ,  $\eta(t-h_2)$  and  $\xi(t+\theta)$  ( $-h \leq \theta < 0$ ) by  $y(t)$ , and  $x(-h_1)$ ,  $\eta(-h_2)$ ,  $\xi(\theta)$  ( $-h \leq \theta < 0$ ) by respectively  $y$ . The quantity  $y(t)$  makes it possible to compose a probability description of the plant  $A$  at the instant  $t$ . The quantities  $J_T$  (23) and  $J_\infty$  (24) with the chosen law of control  $\xi$  may be regarded as functionals with respect to  $y(t_0)$ , that is,

$$M \left\{ \int_{t_0}^T \omega[t, x(t), \xi(t)] dt + \psi(T) \right\} = J_T[t_0, y^0(t_0), \xi] \quad (25)$$

$$\lim_{T \rightarrow \infty} \frac{J_T}{T - t_0} = J_\infty[t_0, y^0(t_0), \xi] \quad (26)$$

It is therefore reasonable in this case to seek the optimal control signal  $\xi^0$  as a function of  $y(t)$ , that is, in the form of a functional

$$\xi(t) = \xi[t, y(t)] \quad (27)$$

Call the admissible control signals the set of such functionals, sufficiently regular to give a meaning to the solution of (21) with  $\xi(t)$  of (27), and, possibly, constrained by supplementary restrictions arising from the statement of the problem (for instance,  $|\xi| \leq 1$ ). Designate the set of admissible control signals by the symbol  $\Xi$ . Now the problem can be formulated.

**Problem 4.** It is required to find a control signal  $\xi^0$  belonging to  $\Xi$  which minimizes (25) for all  $y^0$  belonging to  $Y_0$ ,  $t_0 \geq 0$ .

**Problem 5.** It is required to find a control signal  $\xi^0$  belonging to  $\Xi$  minimizing (26) for all  $y^0$  belonging to  $Y_0$ ,  $t_0 \geq 0$ . Here  $Y_0$  is some region of the components  $y$  given in advance.

Denote by  $x(t, y^0(t_0), \xi)$  the random motion of the system, generated by the initial conditions  $y^0(t_0)$  with a certain choice of the control law; moreover, assume necessarily, with  $t_0 - h \leq t < t_0$ , that the control signal  $\xi(t)$  tallies with that  $\xi(t_0 + \theta)$  ( $t_0 + \theta = t$ ) which is a component of  $y^0(t_0)$ .

Now formulate the criterion of optimality for Problem 4.

**Theorem 3.** It is assumed that for all  $y^0(t_0)$  belonging to  $Y_0$  and  $0 \leq t_0 \leq T$  there exists an admissible control signal  $\xi(t)$  (or  $\xi = \xi[t, y(t)]$ ) such that (25) has a meaning, is finite, and almost all the realizations  $\{x(t, y^0(t_0), \xi), \eta(t, y^0(t_0)), \xi(t + \theta) (-h \leq \theta < 0)\}$  belong to  $Y$ , where  $Y$  is a certain region of values of  $y$ . Let it be possible to find functionals  $v[t, y]$  and  $\xi^0[t, y]$  satisfying the conditions

$$(1) \quad v[T, y(T)] = M\{\psi[x(T, y(T), \xi)]\} \quad (28)$$

for all  $y(T)$  belonging to  $Y$

$$(2) \quad \left( \frac{dM\{v\}}{dt} \right)_{\xi^0} + M\{\omega[t, x(t, y(t), \xi^0), \xi^0]\} \\ = \min_{\xi \in \Xi} \left[ \left( \frac{dM\{v\}}{dt} \right)_{\xi} + M\{\omega[t, x(t, y(t), \xi), \xi]\} \right] = 0 \quad (29)$$

for all  $y(t)$  belonging to  $Y$  and all  $t$  in the closed interval  $[0, T]$ .

Then  $\xi^0[t, y(t)]$  is the optimal control signal for Problem 4 and  $v[t_0, y^0(t_0)] = \min J_T[t_0, y^0(t_0), \xi]$ .

The solution of Problem 5 is obtained by passage to the limit from the solution of Problem 4.

The results of applying the given criterion to a system described by equations of an actual form are illustrated. Consider Problem 5 for the system

$$\frac{dx_i}{dt} = \sum_{j=1}^n a_{ij}(t)x_j(t) + b_i\xi + a_i\eta(t) + \phi \quad (30)$$

with the condition of minimum (26), where

$$J_T = M \left\{ \int_{t_0}^T \left[ \sum_{i,j=1}^n \omega_{ij}(t)x_i(t)x_j(t) + \lambda\xi^2(t) \right] dt \right. \\ \left. + \sum_{i,j=1}^n \psi_{ij}x_i(T)x_j(T) \right\} \quad (31)$$

The delays along both channels 1 and 2 are assumed to be equal to  $h > 0$ , and it is admitted that any initial deviations  $x^0(t_0 - h)$  and  $\eta(t_0 - h)$  belong to  $(\eta_1, \eta_2)$ .

With sufficiently wide assumptions concerning the character of the Markov probability process  $\eta(t)$  and with the condition of full controllability of the system  $dx_i/dt = \sum a_{ij}x_j + b_i\xi$ ,

the functionals  $v[t, y]$  and  $\xi^0[t, y]$  satisfying criterion (21) can be found, and passage to the limit with  $T \rightarrow \infty$  can be carried out. Problems 4 and 5 can also be solved. In addition the following result is valid.

## Results

The optimal control signal for Problems 4 and 5 stated with conditions (30) and (31) has the form

$$\xi^0[t, y(t)] = \sum_{i=1}^n \mu_i(t)x_i(t-h) + v[t, \eta(t-h)] \\ + \int_{-h}^0 \varrho[t, \theta]\xi[t+\theta]d\theta \quad (32)$$

The term  $v$  is determined at every instant  $t$  with respect to the realized  $\eta(t-h)$ , but to calculate it one must know the prediction  $M\{\eta(\tau)/\eta(t-h)\}$  with  $\tau > t-h$ .

Here the functional  $v[t, y(t)]$  has the form of the sum of the quadratic and linear functionals of  $x_i(t-h)$  and  $\xi(t+\theta)$  with coefficients dependent on  $\eta(t-h)$ .

Analysing the resulting solution  $\xi^0$  the following conclusion is arrived at: the optimal control signal  $\xi^0$  chosen here at every instant  $t$  is the same as would be obtained in a deterministic system and without delay of the feedback signals; however here, instead of the known quantities  $x_i(t)$  of the deterministic system, their best mean square predictions  $M\{x_i(t)/x(t-h), \eta(t-h), \xi(t+\theta) (-h \leq \theta < 0)\}$  must enter into the control law, and the deterministic load  $\eta(\tau)$  ( $\tau > t-h$ ) is likewise replaced by the mean prediction  $M\{\eta(\tau)/\eta(t-h)\}$ .

**Case III.** This case reduces naturally to the previous one, and it is not considered individually.

When several of the cases analysed are combined in one system, the statements of the problems, criteria of optimality and results are combined correspondingly.

In conclusion it is observed that Case II can be included in the more general case when incomplete information is transmitted along the feedback channels 1 and 2. For it can be assumed indeed that at the instant  $t$  there are applied to the regulator  $B$  signals  $y(t)$  and  $\zeta(t)$ , statistically connected with  $x(t)$  and  $\eta(t)$  (in Case II,  $\{y(t), \zeta(t)\} = \{x(t-h), \eta(t-h), \xi(t+\theta)\}$ ) and an optimal control signal depending on these signals can be constructed. The foregoing reasoning and conclusions are generalized to this more general case. The quality of the process depends on how much the processes  $\{y(t), \zeta(t)\}$  and  $\{x(t), \eta(t)\}$  are connected informationally, or, in other words, how far the processes  $\{x(t), \eta(t)\}$  are observable<sup>12</sup> with respect to  $\{y(t), \zeta(t)\}$ .

## References

1. FELDBAUM, A. A. Optimal processes in automatic control systems. *Automat. Telemekh.* 14, No. 6 (1953)
2. PONTRYAGIN, L. S., BOLTYANSKII, V. G., GAMKRELIDZE, R. V., and MISHCHENKO, E. F. A mathematical theory of optimal processes. Fizmatgiz (1961)
3. ROZONOER, L. I. Pontryagin's Maximum Principle in the theory of optimal systems. *Automat. Telemekh.* 20, Nos. 10-12 (1959)
4. BELLMAN, R. *Dynamic Programming*. I.I.L. (1960)
5. LERNER, A. YA. Maximum high speed of automatic control systems. *Automat. Telemekh.* 15, No. 6 (1956)

- <sup>6</sup> LERNER, A. YA. *Design Principles for High-speed Following Systems and Regulators*. 1961. Moscow; Gosenergoizdat
- <sup>7</sup> BELLMAN, R., GLICKSBERG, J., and GROSS, O. *Some Aspects of the Mathematical Theory of Control Processes*. 1958. Project Rand
- <sup>8</sup> FELDBAUM, A. A. Calculating devices in automatic systems. *Fizmatgiz* (1959)
- <sup>9</sup> FILIPPOV, A. F. Some questions of the theory of optimal control. *Vestn. MGU* No. 2 (1959)
- <sup>10</sup> KALMAN, R. E., and BERTRAM, J. E. Control systems analysis and design via the 'Second Method' of Liapunov. *Pap. Amer. Soc. Mech. Eng.*, No. 2 (1959)
- <sup>11</sup> KALMAN, R. E. On the general theory of control systems. *Automatic and Remote Control*. 1961, Vol. 2. London; Butterworths
- <sup>12</sup> KALMAN, R. E. New methods and results in linear prediction and filtering theory. *RJAS Tech. Rep.* 61-1 (1961)
- <sup>13</sup> KULIKOVSKI, R. A. *Bull. Acad. Polon. Sci., Serie des sciences techniques*, Vol. VII, No. 6, 11, 12 (1959), Vol. VIII, No. 4 (1960)
- <sup>14</sup> LA SALLE, J. Time optimal control systems. *Proc. nat. Acad. Sci.*, Vol. 45, No. 4 (1959)
- <sup>15</sup> GIRSANOV, I. V. Minimax problems in the theory of diffusion processes. *Dokl. AN SSSR*, Vol. 136, No. 4 (1960)
- <sup>16</sup> TSYPKIN, YA. Z. On optimal processes in pulsed automatic systems. *Dokl. AN SSSR*, Vol. 136, No. 2 (1960)
- <sup>17</sup> BELLMAN, R., and KALABA, R. Theory of dynamic programming and control systems with feedback. *Automatic and Remote Control*. 1961, Vol. 1. London; Butterworths
- <sup>18</sup> MERRIAM, K. U. Calculations connected with one class of optimal control systems. *Automatic and Remote Control*. 1961, Vol. 3. London; Butterworths
- <sup>19</sup> BUTKOVSKII, A. G., and LERNER, A. YA. Optimal control of systems with distributed parameters. *Dokl. AN SSSR*, Vol. 134, No. 4 (1960)
- <sup>20</sup> KRAMER, J. On control of linear systems with time lags. *Inform. Control*, Vol. 3, No. 4 (1960)
- <sup>21</sup> KHARATISHVILI, G. L. The maximum principle in the theory of optimal processes with time lags. *Dokl. AN SSSR*, Vol. 136, No. 1 (1961)
- <sup>22</sup> BELLMAN, R., and KALABA, R. Dynamic programming and control processes. *J. Bas. Engng.* (March 1961)
- <sup>23</sup> LETOV, A. M. Analytical construction of regulators. *Automat. Telemekh.*, Vol. 21, Nos. 4-6 (1960)
- <sup>24</sup> LETOV, A. M. Analytic construction of regulators; the dynamic programming method. *Automat. Telemekh.*, Vol. 22, No. 4 (1961)
- <sup>25</sup> FELDBAUM, A. A. Information storage in closed systems of automatic control. *Izv. AN SSSR, Otdelenie tekhnicheskikh nauk. Energet-automat.*, No. 4 (1961)
- <sup>26</sup> BELLMAN, R. Adaptive control processes. 1961. *Project Rand*
- <sup>27</sup> LIAPUNOV, A. M. *The General Theory of Stability of Motion*. 1950. Gostekhizdat
- <sup>28</sup> CHETAEV, N. G. *Stability of Motion*. 1956. Gostekhizdat
- <sup>29</sup> KRASOVSKII, N. N. *Some Problems of the Theory of Stability of Motion*. 1960. Gostekhizdat
- <sup>30</sup> KRASOVSKII, N. N., and LIDSKII, E. A. Analytic construction of regulators in systems with random parameters. *Automat. Telemekh.* Vol. 22, Nos. 9-11 (1961)
- <sup>31</sup> KRASOVSKII, N. N. A problem of tracking. *Prikl. matemat. mekh.*, Vol. 26, No. 2 (1962)
- <sup>32</sup> KRASOVSKII, N. N. Analytic construction of an optimal regulator in a system with time lags. *Prikl. matemat. mekh.* Vol. 26, No. 1 (1962)

# Optimal Control of Systems with Distributed Parameters

A. G. BUTKOVSKII

## Summary

Plants with distributed parameters are widespread in many branches of engineering. The problem of designing control systems which are the best in a given sense for such plants requires the development of new techniques of optimal control. Control systems with distributed parameters have a number of specific properties that make investigation into them very complex compared with lumped-constant systems. This paper gives a statement of optimal control problems on systems with distributed parameters for a fairly broad class of systems described by non-linear integral equations under arbitrary constraints. It formulates a maximum principle for this case, which gives the necessary conditions for determining optimal controls.

The paper also considers approximate methods of solving optimal control problems on systems with distributed parameters, based on approximating the equations for the distribution functions by ordinary differential equations.

## Sommaire

Dans de nombreux domaines de la technique on rencontre des installations avec des paramètres répartis. Le problème de l'étude des systèmes de réglage de telles installations qui seraient les meilleurs dans un certain sens, nécessite le développement de nouvelles méthodes d'optimisation. En effet, les systèmes de réglage avec des paramètres répartis ont certaines propriétés qui rendent leur investigation plus complexe que celle de systèmes de réglage avec des paramètres concentrés. Le rapport montre comment traiter l'optimisation de tels systèmes décrits au moyen d'équations intégrales non-linéaires et certaines contraintes arbitraires. Il formule un principe de maximisation qui donne les conditions de détermination de système d'optimisation.

En outre, ce rapport considère certaines méthodes approchées d'optimisation de systèmes avec des paramètres répartis, méthodes basées sur une approximation des fonctions de distribution au moyen d'équations différentielles ordinaires.

## Zusammenfassung

In vielen Zweigen der Technik finden sich Anlagen mit verteilten Parametern. Die Schwierigkeit, irgendwie optimale Regelungen für solche Anlagen zu entwerfen, macht die Entwicklung neuer Verfahren zur Optimalwertregelung notwendig. Regelsysteme mit verteilten Parametern haben eine Anzahl besonderer Eigenschaften, die ihre Untersuchung — verglichen mit der Untersuchung von Systemen mit zusammengefaßten Konstanten — erschweren. In diesem Beitrag werden die Probleme der Optimalregelung von Systemen mit verteilten Parametern formuliert. Diese Formulierung gilt für eine große Anzahl von Systemen, die durch nicht-lineare Integralgleichungen unter beliebigen Grenzbedingungen beschrieben sind. Für diesen Fall wird ein Maximumprinzip angegeben, das auch die notwendigen Bedingungen für die Bestimmung der Optimalregelung enthält.

In diesem Beitrag werden auch Näherungsmethoden für die Lösung von Optimalregelungsproblemen bei Systemen mit verteilten Parametern betrachtet, die auf einer Näherung der Verteilungsfunktionen mittels gewöhnlicher Differentialgleichungen beruhen.

In many engineering applications the need arises for control of systems with parameters that are distributed in space. A wide class of industrial and non-industrial processes falls within this

category: production flow processes, heating of metal in methodical or straight-through furnaces before rolling or during heat-treatment, establishment of given temperature distributions in 'thick' ingots, growing of monocrystals, drying and calcining of powdered materials, sintering, distillation, etc., right through to the control of the weather.

The processes in such systems are normally described by partial differential equations, integral equations, integro-differential equations, etc.

The problem of obtaining the best operating conditions for the installation (the highest productivity, minimum expenditure of raw material and energy, etc.) under given additional constraints has required the development of an appropriate mathematical apparatus capable of determining the optimal control actions for the plant.

Pontryagin's maximum principle and Bellman's dynamic programming method have been the most interesting results in this direction for systems with lumped parameters.

A wide class of systems with distributed parameters is described by a non-linear integral equation of the following form:

$$Q(P) = \int_D K[P, S, Q(S), U(S)] dS \quad (1)$$

Here the matrix

$$Q(P) = \begin{pmatrix} Q^1(P) \\ \vdots \\ Q^n(P) \end{pmatrix} = \|Q^i(P)\| \quad (2)$$

describes the condition of the controlled system with distributed parameters, while the matrix

$$U(P) = \begin{pmatrix} U^1(P) \\ \vdots \\ U^r(P) \end{pmatrix} = \|U^i(P)\| \quad (3)$$

describes the control actions on the system. Here and in the following, the index  $i$  will refer to a row number and  $j$  to a column number in a matrix. The point  $P$  belongs to a certain fixed  $m$  dimensional region  $D$  in Euclidean space.

The components of the single-column matrix

$$K(P, S, Q, U) = \begin{pmatrix} K^1(P, S, Q, U) \\ \vdots \\ K^n(P, S, Q, U) \end{pmatrix} = \|K^i(P, S, Q, U)\| \quad (4)$$

belong to class  $L_2$  and have continuous partial derivatives w.r.t. the components of the matrix  $Q$ .

It will be assumed that the function  $U(P)$  is piecewise discontinuous, its values being chosen from a certain fixed permissible set  $\Omega$ . Controls  $U(P)$  having this property will be called permissible.

Further, from the set of conditions  $Q(P)$  and controls  $U(P)$ , related by integral eqn (1), let  $q$  functionals be determined, having a continuous gradient (weak Gato differential).

$$I^i = I^i[Q(P)], \quad i=0, 1, \dots, l \quad (5)$$

$$I^i = I^i[Q(P), U(P)] = \Phi^i(z), \quad i=l+1, \dots, q \quad (6)$$

where

$$z = \begin{pmatrix} z^0 \\ \vdots \\ z^k \end{pmatrix} = \begin{pmatrix} \int_D F^0[S, Q(S), U(S)] dS \\ \vdots \\ \int_D F^k[S, Q(S), U(S)] dS \end{pmatrix} = \begin{pmatrix} \int_D F[S, Q(S), U(S)] dS \end{pmatrix} \quad (7)$$

The functions  $\Phi^i(z)$ ,  $i=l+1, \dots, q$  and  $F^i(S, Q, U)$ ,  $i=0, 1, \dots, k$ , are continuous and have continuous partial derivatives w.r.t. the components of the matrices  $z$  and  $Q$  respectively.

The optimal control problem is formulated in the following manner.

It is required to find a permissible control  $U(P)$  such that by virtue of equation (1)

$$I^i = 0, \quad i=0, 1, \dots, p-1, p+1, \dots, q \quad (8)$$

while the functional  $I^p$  assumes its smallest value. Here  $p$  is a fixed index,  $0 \leq p \leq q$ .

The following rectangular matrices are introduced

$$\frac{\partial \Phi}{\partial z} = \left\| \frac{\partial \Phi^i}{\partial z^j} \right\|; \quad i=0, 1, \dots, l; \quad j=0, 1, \dots, k \quad (9)$$

$$\frac{\partial F}{\partial Q} = \left\| \frac{\partial F^i}{\partial Q^j} \right\|; \quad i=0, 1, \dots, k; \quad j=1, 2, \dots, n \quad (10)$$

$$\text{grad } I = \|\text{grad}_j I^i\|; \quad i=l+1, \dots, q; \quad j=1, 2, \dots, n \quad (11)$$

where  $\text{grad}_j I^i$  denotes the  $j$ th component of the vector  $\text{grad } I^i$  w.r.t. the coordinate  $Q^j$ .

The following theorem<sup>5</sup> can be used as the basis of a solution of the problem formulated above on the optimum control of a plant with distributed parameters.

**Theorem.** Let  $U = U(S)$  be a permissible control such that by virtue of eqn (1) the conditions (8) are satisfied and the matrix function  $M(P, R) = \|M_{ij}(P, R)\|$ ,  $i, j = 1, 2, \dots, n$ , satisfies the integral equation [linear in  $M(P, R)$ ]

$$\begin{aligned} M(P, R) + \frac{\partial}{\partial Q} K[P, R, Q(R), U(R)] \\ = \int_D M(P, S) \frac{\partial}{\partial Q} K[S, R, Q(R), U(R)] dS \\ = \int_D \frac{\partial}{\partial Q} K[P, S, Q(S), U(S)] M(S, R) dS \end{aligned} \quad (12)$$

Then for this control,  $U(S)$ , to be optimal there must exist one-row numerical matrices

$$a = \|c_0, c_1, \dots, c_l\| \quad \text{and} \quad b = \|c_{l+1}, \dots, c_q\| \quad (13)$$

of which at least one is not null, and also  $c_p \leq 0$ , such that for almost all fixed values of the argument  $S \in D$  the function

$$\begin{aligned} \pi(S, U) = & a [\text{grad } I\{Q(P)\}, K\{P, S, Q(S), U\} \\ & - \int_D M(P, R) K\{R, S, Q(S), U\} dR] \\ & + b \frac{\partial}{\partial z} \Phi \left[ \int_D F\{P, Q(P), U(P)\} dP \right] \\ & \cdot \left[ \frac{\partial}{\partial Q} F\{P, Q(P), U(P)\}, K\{P, S, Q(S), U\} \right. \\ & \left. - \int_D M(P, R) K\{R, S, Q(S), U\} dR \right] \\ & + b \frac{\partial}{\partial z} \Phi \left[ \int_D F\{P, Q(P), U(P)\} dP \right] \cdot F\{S, Q(S), U\} \end{aligned} \quad (14)$$

of the variable  $U \in \Omega$  attains a maximum, i.e. for almost all  $S \in D$  the following relation holds:

$$\pi(S, U) = H(S) \quad (15)$$

where

$$H(S) = \sup_{u \in \Omega} \pi(S, U) \quad (16)$$

As an example of the application of this theorem, consider the important practical problem of the heating of a massive body in a furnace. Let the temperature distribution along the  $x$  axis,  $0 \leq x \leq L$ , at any instant  $t$ ,  $0 \leq t \leq T$ , be described by the function  $Q = Q(x, t)$ . Here the temperature  $U(t)$  of the heating medium, which in this case is the controlling agent, is a function constrained by the conditions

$$A_1 \leq U(t) \leq A_2, \quad 0 \leq t \leq T \quad (17)$$

i.e. in this case the set  $\Omega$  is the interval  $[A_1, A_2]$ .

It is known that the distribution function  $Q(x, t)$ , if initially zero, is related to the control  $U(t)$  by the following integral equation

$$Q(x, t) = \int_0^t K(x, t, \tau) U(\tau) d\tau \quad (18)$$

where  $K(x, t, \tau)$  is a known weighting function.

In the heating of a body there is usually given a temperature distribution  $Q^* = Q^*(x)$  which is required to be attained in the minimum time. However, if the equation

$$Q(x, t) = Q^*(x) \quad (19)$$

for any permissible control is not satisfied for any fixed  $t$ ,  $0 \leq t \leq T$ , then the problem becomes that of determining a permissible control  $u(t)$ ,  $0 \leq t \leq T$ , such that the functional

$$I^0 = \int_0^L [Q^*(x) - Q(x, T)]^\gamma dx \quad (20)$$

attains its minimum. Here  $\gamma$  is a positive even integer.

Since the integrand in eqn (18) is independent of the controlled function  $Q(x, t)$ , then according to eqn (12) the function  $M(t, c) \equiv 0$  for all  $t$  and  $\tau$  in the interval  $[0, T]$ .

It follows that the function  $\pi(\tau, U)$  takes the form

$$\begin{aligned}\pi(\tau, U) &= c_0 \int_0^L \frac{\partial}{\partial Q} [Q^*(x) - Q]^\gamma \cdot K(x, T, \tau) U dx \\ &= -\gamma c_0 U \int_0^L [Q^*(x) - Q(x, T)]^{\gamma-1} K(x, T, \tau) dx\end{aligned}\quad (21)$$

Since in this case by the conditions of the theorem  $c^0 < 0$ , so  $-\gamma c_0 > 0$ , and hence the maximum of  $\pi(\tau, U)$  w.r.t.  $U$ , with  $A_1 \leq U \leq A_2$ , is reached when

$$\begin{aligned}U(\tau) &= \frac{A_1 + A_2}{2} \\ &+ \frac{A_2 - A_1}{2} \operatorname{sgn} \int_0^L [Q^*(x) - Q(x, T)]^{\gamma-1} K(x, T, \tau) dx\end{aligned}\quad (22)$$

If we substitute expression (18) for  $Q(x, t)$  in eqn (22), then we obtain an integral equation for determining the optimum control action  $U(\tau)$ .

For example, if  $\gamma = 2$ ,  $A_1 = -1$ ,  $A_2 = 1$ , then the optimum control action satisfies the following integral equation:

$$U(\tau) = \operatorname{sgn} \left[ Q^*(x) - \int_0^T K(x, T, \tau) U(\tau) d\tau \right] K(x, T, \tau) dx \quad (23)$$

Opening the brackets and altering the order of integration, one finally gets

$$U(\tau) = \operatorname{sgn} \left[ B(\tau) - \int_0^T N(\tau, \theta) U(\theta) d\theta \right] \quad (24)$$

where  $N(\tau, \theta)$  is the symmetrical nucleus

$$N(\tau, \theta) = \int_0^L K(x, T, \tau) K(x, T, \theta) dx \quad (25)$$

$$B(\tau) = \int_0^L Q^*(x) K(x, T, \tau) dx \quad (26)$$

Methods of approximating partial differential equations by finite difference equations can be applied successfully to the approximate solution of problems of the optimal control of systems with distributed parameters. This has the advantage that results obtained for lumped-parameter optimal systems can be used.

As an example, consider the optimal control of a system described by the following equation

$$\frac{\partial Q}{\partial t} = a \frac{\partial^2 Q}{\partial x^2}, \quad Q = Q(x, t), \quad 0 \leq x \leq S, \quad 0 \leq t \leq T \quad (27)$$

with these initial and boundary conditions

$$Q(x, 0) = Q_0(x) \quad (28)$$

$$\left. \frac{\partial Q}{\partial x} \right|_{x=0} = \alpha [U(t) - Q(0, t)], \quad \left. \frac{\partial Q}{\partial x} \right|_{x=S} = 0 \quad (29)$$

Also let the function  $Q^* = Q^*(x)$  be given. The problem may be formulated in double form:

(a) To find a permissible control  $U(t)$ ,  $0 \leq t \leq T$ ,  $U \in \Omega$  ( $\Omega$  is the set of permissible control values), such that the equation

$$Q(x, T) = Q^*(x), \quad 0 \leq x \leq S \quad (30)$$

is satisfied for a minimal time  $T$ .

However, in many cases eqn (30) cannot be accurately satisfied for any  $T$ . It then makes sense to formulate the problem as follows:

(b) To find a permissible control  $U(t)$ ,  $U \in \Omega$ ,  $0 \leq t \leq T$ , where  $T$  is a fixed time, such that the functional

$$I = \int_0^S [Q^*(x) - Q(x, T)]^\gamma dx \quad (31)$$

which characterizes the measure of deviation of the actual distribution from the given one ( $\gamma$  a positive even integer), should reach a minimum.

Using the straight-line method, problems (a) and (b) may be reduced to an ordinary problem of optimum control for systems with lumped parameters.

In fact, splitting the interval  $[0, S]$  on the  $x$  axis into  $n$  equal parts by the points  $x_0 = 0$ ,  $x_1 = s$ , ...,  $x_n = S$ , where  $s = S/n$ , and replacing the second partial derivative of  $Q(x, t)$  w.r.t.  $x$  in eqn (27) by the second difference ratio, we obtain a finite system of order  $(n+1)$  of ordinary linear differential equations for the functions  $q_i(t)$ ,  $i = 0, 1, \dots, n$ :

$$\begin{aligned}\dot{q}_0 &= -(\sigma + \beta) q_0 + \sigma q_1 + \beta U \\ \dot{q}_i &= \sigma (q_{i-1} - 2q_i + q_{i+1}), \quad i = 1, 2, \dots, n-1 \\ \dot{q}_n &= \sigma (q_{n-1} - q_n)\end{aligned}\quad (32)$$

with the initial condition

$$q_i(0) = Q_0(is), \quad i = 0, 1, \dots, n \quad (33)$$

and the final condition

$$q_i(T) = Q^*(is), \quad i = 0, 1, \dots, n \quad (34)$$

Here  $\beta$  and  $\sigma$  are constant coefficients which can be expressed in terms of  $a$  and  $\alpha$ .

In problem (a) the functional that has to be minimized is the time  $T$ . This problem can be solved by using the maximum principle. Gamkrelidze<sup>3</sup> has shown that its solution always exists and is unique.

In problem (b) the functional to be minimized is

$$I = \sum_{i=1}^n [q_i(T) - Q^*(is)]^\gamma \quad (35)$$

In certain cases it is required to determine the optimum variation law for a control action which is itself distributed in space, constraints being placed on it in time and also in space coordinates.

For example, it is sometimes material that too great space variations cannot be allowed in certain physical quantities such as temperature, pressure, electric field, etc.

We shall consider as an illustration the heat-exchange equation

$$b(y, t) \frac{\partial}{\partial t} Q(y, t) + b(y, t) v(t) \frac{\partial}{\partial y} Q(y, t) + Q(y, t) = U(y, t) \quad (36)$$

for the exchange between a stationary heating medium with temperature  $U = U(y, t)$ ,  $0 \leq y \leq L$ ,  $0 \leq t \leq T$  ( $y$  being a space coordinate and  $t$  the time), and a material moving at velocity  $v = v(t) \geq 0$  in the positive sense along the  $y$  axis and becoming heated in the process of moving over the interval  $0 \leq y \leq L$ . The state of heating of the material is described by the function  $Q(y, t)$ . The initial and boundary conditions take the form

$$Q(y, 0) = Q_0(y), \quad Q(0, t) = 0 \quad (37)$$

In this case a permissible control is considered to be a function  $U = U(y, t)$ ,  $0 \leq y \leq L$ ,  $0 \leq t \leq T$ , that satisfies the conditions

$$A_1 \leq U(y, t) \leq A_2 \quad (38)$$

$$A_3 \leq \frac{\partial}{\partial y} U(y, t) \leq A_4 \quad (39)$$

where  $A_1, A_2, A_3$  and  $A_4$  are given constants.

Physically these constraints correspond to the fact that in feed-through heating installations one cannot allow too great amplitudes of temperature fluctuation in the heating medium, or excessive temperature drop over the length of the furnace.

In this case one has to determine the control  $U = U(y, t)$ , subject to conditions (38) and (39), such that, in spite of all possible disturbances of the heating process caused by variations in the velocity  $v(t)$  and by variations in the thermal parameters  $b(y, t)$  of the process, the deviation of the temperature of the material leaving the furnace from a certain given temperature  $Q^*$  should be on the average a minimum, i.e. one has to minimize the functional

$$I = \int_0^T [Q^* - Q(L, t)]^\gamma dt \quad (40)$$

where  $\gamma$  is a positive even integer.

In order to reduce the partial differential equations to difference differential equations, split up the interval  $[0, L]$  on the  $y$  axis into  $n$  equal parts by the points  $y_0 = 0, y_1 = 1, \dots, y_n = L$ , where  $l = L/n$ . Replacing the partial derivative w.r.t.  $y$  in eqn (36) by the difference ratio, we obtain a system of order  $n$  of ordinary linear differential equations in the functions  $q_i(t)$ ,  $i = 1, 2, \dots, n$ ,

$$b(il, t) \dot{q}_i + \frac{1}{l} b(il, t) v(t) [q_i - q_{i-1}] + q_i = U_i(t) \quad (41)$$

with  $q_0(t) = 0$ ,  $0 \leq t \leq T$  and  $q_i(0) = Q_0(il)$ .

Equation (41) may be rewritten in the form

$$\dot{q}_i = \beta q_{i-1} + \alpha_i q_i + U_i, \quad i = 1, 2, \dots, n \quad (42)$$

where  $U_i = U_i(t) = U(il, t)$ , while the coefficients  $\beta$  and  $\alpha_i$  can be expressed explicitly in terms of the functions  $v(t)$  and  $b(il, t)$ .

According to conditions (38) and (39) the function  $U_i(t)$ ,  $0 \leq t \leq T$ , is subject to the constraints

$$A_1 \leq U_i(t) \leq A_2, \quad i = 1, 2, \dots, n \quad (43)$$

$$A_3 l \leq U_{i+1}(t) - U_i(t) \leq A_4 l, \quad i = 1, 2, \dots, n-1 \quad (44)$$

The functional (40) must now be replaced by:

$$I = \int_0^T [Q^* - q_n(t)]^\gamma dt \quad (45)$$

It can now already be seen that the maximum principle may be used for determining the optimum control actions  $U_i(t)$ ,  $i = 1, 2, \dots, n$ ,  $0 \leq t \leq T$ . In this case the permissible region  $\Omega$  from which the values of the control vector  $U(t) = U_1(t), \dots, U_n(t)$  may be chosen is a closed convex polyhedron in  $n$  dimensional space, described by eqns (43) and (44).

Observe that the function  $H(\psi_i, U_i)$  which has to be maximized according to the maximum principle in this case for each fixed  $t$ ,  $0 \leq t \leq T$ , takes the linear form in the  $U_i$

$$H(\psi_i, U_i) = \sum_{i=1}^n \psi_i U_i \quad (46)$$

Hence the problem of determining the optimum control actions  $U_i(t)$ ,  $i = 1, 2, \dots, n$ , at any instant  $t$ ,  $0 \leq t \leq T$ , reduces to the linear programming problem of maximizing the function  $H$  while satisfying conditions (43) and (44).

Therefore, besides the accurate and quite general methods of solving optimal control problems which have a wide application in engineering, great significance is also attached to approximate methods of solving these problems, based on approximating partial differential equations by ordinary differential equations.

## References

- BOLTYANSKIY, V. G., GAMKRELIDZE, R. V., and PONTRYAGIN, L. S. The theory of optimal processes. *Izv. Akad. Nauk SSSR, Seriya Matematicheskaya* 24, No. 1 (1960)
- BUTKOVSKY, A. G., and LERNER, A. YA. The optimal control of systems with distributed parameters. *Automat. Telemekh.* 21, No. 6 (1960)
- GAMKRELIDZE, R. V. The theory of processes in linear systems that are optimal for rapid response. *Izv. Akad. Nauk SSSR, Seriya Matematicheskaya* 24 (1960)
- BUTKOVSKY, A. G. Optimal processes in systems with distributed parameters. *Automat. Telemekh.* 22, No. 1 (1961)
- BUTKOVSKY, A. G. The maximum principle for optimal systems with distributed parameters. *Automat. Telemekh.* 22, No. 10 (1961)
- BUTKOVSKY, A. G. Some approximate methods for solving optimal control problems on systems with distributed parameters. *Automat. Telemekh.* 22, No. 12 (1961)
- PONTRYAGIN, L. S., BOLTYANSKIY, V. G., GAMKRELIDZE, R. V., MISHCHENKO, YE. F. *The Mathematical Theory of Optimal Processes*. 1961. Fizmatgiz

## DISCUSSION

I. McCausland, *Churchill College, University of Cambridge, England*

My question concerns the method of dealing with the end effects when a distributed-parameter system represented by eqns (27) and (29) is approximated by a lumped-parameter system represented by eqn (32).

Consider a typical system represented by eqns (27) and (29), namely an electrical transmission line of length  $S$  with distributed series resistance and shunt capacitance, with the end at  $x = S$  open-circuited. The voltage between the conductors is  $Q(x, t)$ , and the end at  $x = 0$  is connected to a generator of voltage  $U(t)$  through a resistance  $1/\alpha$ .



If the transmission line is subdivided into  $n$  sections in such a way that the voltages at  $x = 0, x = S/n, \dots, x = S$  are explicitly represented, it seems that each section should be represented by a lumped-parameter approximation as shown in Figure A, where  $R$  and  $C$  are the total resistance and capacitance of the line. If these  $n$  sections are connected together, the complete system is as shown in Figure B (including the input resistance  $1/\alpha$ ).

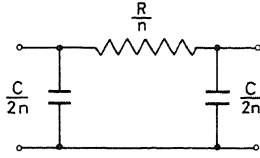


Figure A

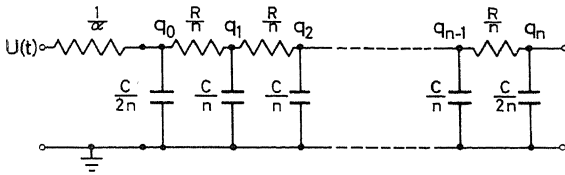


Figure B

It will be seen in Figure B that each end capacitor is half the capacitance of each intermediate capacitor. For this reason the first and last differential equations of this system are different from those of eqns (32) of the paper. For example,  $\dot{q}_n$  is double the value given by eqns (32), and  $\dot{q}_0$  is also different. In view of these differences, I would be grateful for the author's comments on the suitability of different methods of subdivision.

#### A. G. BUTKOVSKII, in reply

In deriving the finite difference approximations of eqns (27) and (29) in the paper I proceeded from a *thermal* interpretation of these equations. The finite-difference system (32) was obtained by dividing a solid of thickness  $S$  into  $n$  equal layers, which were roughly considered to be thin bodies, in the thermal-engineering sense of the word.

Compiling the elementary thermal balances of each elementary layer gives eqn (32). In the process, for the first (far left-hand) layer, account is taken of its interaction with the medium in accordance with the law of convective heat exchange. (The same is done for the last, far right-hand layer, which is simply insulated.)

However, the difference scheme can, of course, be obtained by purely formal means from eqns (27) and (29), by breaking down the segment  $[0, S]$  of the  $x$  axis, for example, into  $n + 2$  equal intervals by points  $x_0 = 0, x_1 = s, x_2 = 2s, \dots, x_{n+2} = S$  where  $s = S/(n + 2)$ .

Assuming that within each  $i$ th interval the temperature of the body is defined by the magnitude  $q_i, i = 0, 1, \dots, n + 1$ , and replacing the partial derivative in eqn (29) by the difference ratio and in eqn (27) the second partial derivative by the second difference ratio, we obtain

$$-\lambda \frac{q_1 - q_0}{s} = \alpha(u - q_0)$$

$$q_{n+1} - q_n = 0$$

$$\dot{q}_i = \frac{a}{s^2} (q_{i-1} - 2q_i + q_{i+1}), i = 1, 2, \dots, n$$

Expressing  $q_0$  from the first equation and substituting ( $i = 1$ ) in the third gives the following system in non-dimensional form:

$$\begin{cases} \dot{q}_1 = -\xi q_1 + q_2 + \eta u \\ \dot{q}_i = q_{i-1} - 2q_i + q_{i+1}, i = 2, 3, \dots, n-1 \\ \dot{q}_n = q_{n-1} - q_n \end{cases}$$

where

$$\dot{q} = \frac{dq}{dt}, \tau \frac{at}{s^2} = (n+2)^2 Fo, \xi = \frac{1+2\beta}{1+\beta}, \eta = \frac{\beta}{1+\beta}, \beta = \frac{\alpha s}{\lambda} = \frac{Bi}{n+2}$$

and  $Fo$  and  $Bi$  are the Fourier and Bio criteria for the given body.

D. P. PETERSEN, *United Aircraft Corporate Systems Center, Farmington, Conn., U.S.A.*

This interesting paper broaches the important question of the control of dynamic systems with distributed parameters. The author lists in his opening paragraph some of the processes in which distributed measurement and control can be expected to be applied. I am engaged in the study and design of systems of weather information processing (not, as yet, of weather control), whence my interest arises.

The use of the word 'control' in the title is perhaps misleading, since the paper deals only with the selection of an optimum input programme, distributed in space and time, to achieve or approximate the desired output (space-time) variation according to certain constraints and criteria. The familiar and critical problems usually associated with control theory (stability, response character and speed, etc.) are thus avoided. Although 'disturbances' to the system are mentioned, it is not clear how these are to be dealt with.

Basic to any concept of control is the provision of means for measurement of one or more critical variables of the process. Without the recurrent flow of information from such measurements, logical operations are meaningless and disturbances and abnormalities cannot be compensated. In uni-dimensional systems (in which time is the only independent variable) and where a finite number of state variables is postulated, measurement feasibility is usually tacitly assumed in the theoretical analysis, and the design of suitable instruments is relegated to the component developers. In a multi-dimensional system, however, this question cannot be glibly side-stepped, since a spatial region cannot even conceptually be filled with an infinite density of measurement apparatus. Furthermore, economic considerations dictate even more sharply for such systems the necessity of seeking efficient dispositions of measurement and control points in the space-time manifold. Thus, in my opinion, analyses based upon an assumption of essentially continuous measurement and control are not likely to lead to useful application. Such systems involve inherently and significantly discrete (multi-dimensional) sampling controllers.

The question of sampling representations of multi-dimensional functions has been studied recently<sup>1, 2</sup>. However, the approach thus far has been in an informational sense rather than from a control point of view.

The following questions are asked:

(1) Has the author investigated any problems of multi-dimensional feedback control?

(2) On what basis should the density of quantization [as discussed after eqn (40)] be chosen in order not to influence adversely the accuracy of solution for the optimum input programme?

(3) Has any apparatus been constructed according to the theory discussed in the paper? If so, does it involve periodic measurement of the output and re-computation of control inputs? What choice was made for the density and frequency of such measurements? Were various arrangements tried, either by simulation or by experiment, and what were the results of such variations?

#### References

- <sup>1</sup> MIYAKAWA, H. *J. Elec. Comm. Engng, Japan*, 42 (1959) 421
- <sup>2</sup> PETERSEN, D. P. and MIDDLETON, D. *Information and Control*, 5 (1962) 279

A. G. BUTKOVSKII, *in reply*

Dr. Petersen is quite right to spotlight the question of obtaining initial information about a distributed-parameter controlled plant. This is, in fact, one of the very difficult and important questions, both in the theory of optimal control of distributed-parameter systems, and in the practical realization of such systems.

I must add that we also are greatly interested in these questions, and have had to face them in creating an optimal control system for a continuous billet-heating furnace<sup>1</sup>. In this case we began by building an adaptive analogue of a distributed-parameter plant<sup>2, 3</sup>.

All the main disturbance effects exerted on the plant were measured; they were fed to an analogue-digital model of the plant. The parameters of the plant were so adjusted as to minimize the deviation of the output parameter of the model from the value of the same parameter, as

measured directly on the plant. After the period of self-adjustment, information on the state of the controlled plant was used in the controller, which realized the optimal control algorithm, found from theoretical examinations.

#### References

- <sup>1</sup> BUTKOVSKII, A. G. and LERNER, A. Ya. Optimal control of distributed-parameter systems, *Automat. telemekh.* 21, No. 6 (1960)
- <sup>2</sup> BUTKOVSKII, A. G. Simulation of some elements with distributed parameters, *Automatic Control Symposium* (1961). Moscow; *Izd. Akad. Nauk SSSR*.
- <sup>3</sup> LERNER, A. YA. Optimal control for continuous processes. *2nd IFAC Congr.*, Basle, 1963. London: Butterworths; Munich: Oldenbourg

# Solution of Optimum Control Problems by using Pontryagin's Maximum Principle

Y. SAKAWA and C. HAYASHI

## Summary

This paper deals with the solution of optimum control problems, using Pontryagin's maximum principle. Since an optimum control system which completes the transient response in minimum time behaves like a relay servo, the important task of the controller is to reverse the sign of the manipulated variable at the proper instants according to the operating condition of the system. A set of points at which the sign of the manipulated variable is to be reversed forms a switching surface in the phase space. In this paper the optimum switching surface of a linear third-order system is investigated. Also dealt with is the optimum switching curve of a non-linear control system governed by van der Pol's equation with an additional forcing term.

## Sommaire

Ce rapport s'occupe de la solution des problèmes de la commande optimale en utilisant le principe du maximum de Pontryagin. Puisqu'un système de commande optimale qui réalise la réponse transitoire dans un minimum de temps fonctionne comme un asservissement à relais, la tâche importante du régulateur est d'inverser le signe de la grandeur réglée aux instants appropriés, selon les conditions de fonctionnement du système. Un ensemble de points, auxquels il faut inverser le signe de la grandeur réglée forme une surface de commutation dans l'espace des phases. Dans ce rapport on recherche la surface optimale de commutation d'un système linéaire du troisième ordre. On traite aussi la courbe optimale de commutation d'un système de commande non linéaire régi par l'équation de van der Pol avec un terme de contrainte supplémentaire.

## Zusammenfassung

Der Beitrag beschäftigt sich mit der Lösung von optimalen Regelproblemen unter Anwendung des Pontryaginschen Maximumprinzips. Da ein optimales Regelsystem, bei dem der Übergangsvorgang in minimaler Zeit abgeschlossen ist, sich im wesentlichen wie ein Relaisstellglied verhält, ist es die Hauptaufgabe des Reglers, das Vorzeichen der Stellgröße im richtigen Zeitpunkt, entsprechend den Arbeitsbedingungen des Systems, umzukehren. Eine Menge von Punkten, in denen das Vorzeichen der Stellgröße umgekehrt wird, bildet innerhalb des Phasenraumes eine Schaltfläche. In dieser Arbeit wird die optimale Schaltfläche eines linearen Systems dritter Ordnung untersucht. Außerdem enthält sie eine Betrachtung über die optimale Schaltkurve eines nichtlinearen Regelsystems, das durch die van der Pol'sche Gleichung mit zusätzlicher Anregungsgröße beschrieben wird.

## Introduction

This paper treats the solution of optimum control problems by using Pontryagin's maximum principle<sup>1</sup>. It is assumed that the magnitude of the manipulated variable is limited by saturation. In order to complete the transient response in minimum time, the magnitude of the manipulated variable is kept at its maximum value throughout the transient. This implies that the control system behaves like a relay servo. The important task of the

controller is to reverse the sign of the manipulated variable at the proper instants according to the operating condition of the system. A set of points at which the sign of the manipulated variable is to be reversed forms a switching surface in the phase space<sup>2, 3</sup>.

In the following, the optimum switching surface of a linear third-order system is investigated and the optimum switching curve of a non-linear control system governed by van der Pol's equation, with additional forcing term, is also dealt with.

## Optimum Control of a Linear Third-order Servo System

The transfer function of this type of system is given by

$$G(s) = \frac{1}{s(s^2 + 2\zeta s + 1)} \quad (1)$$

where  $\zeta$  is a positive constant. It is assumed that the manipulated variable  $u$  of the system is subject to the constraint:

$$|u| \leq 1 \quad (2)$$

The magnitude of the manipulated variable  $u$  must be kept at its maximum value during the control process<sup>1</sup>. Therefore, the controller must be of the relay type. The block diagram of the system is shown in Figure 1. The purpose here is to design the optimum controller which leads the system to a desired equilibrium state in minimum time.

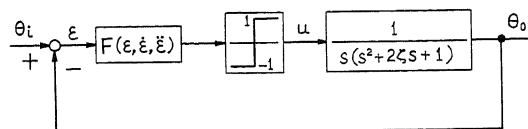


Figure 1. Block diagram of a third-order servo system

Assuming that the input signal  $\theta_i$  is a step function, the differential equation of the system is obtained:

$$\frac{d^3 \epsilon}{dt^3} + 2\zeta \frac{d^2 \epsilon}{dt^2} + \frac{d\epsilon}{dt} + u = 0 \quad (3)$$

where  $\epsilon$  is the error signal. Eqn (3) is equivalent to the following set of linear equations:

$$\left. \begin{aligned} \frac{d\epsilon_1}{dt} &= \epsilon_2 \\ \frac{d\epsilon_2}{dt} &= \epsilon_3 \\ \frac{d\epsilon_3}{dt} &= -\epsilon_2 - 2\zeta \epsilon_3 - u \end{aligned} \right\} \quad (4)$$

where  $\varepsilon_1 = \varepsilon$ . Let the poles of the transfer function (1) be denoted by

$$\lambda_1 = -\zeta + (\zeta^2 - 1)^{\frac{1}{2}}, \lambda_2 = -\zeta - (\zeta^2 - 1)^{\frac{1}{2}}, \lambda_3 = 0 \quad (5)$$

then Lurie's transformation<sup>4</sup> for eqns (4) may be written as (assuming that  $\lambda_1 \neq \lambda_2$ )

$$\left. \begin{aligned} \varepsilon_1 &= \frac{x_1}{\lambda_1(\lambda_2 - \lambda_1)} + \frac{x_2}{\lambda_2(\lambda_1 - \lambda_2)} - x_3 \\ \varepsilon_2 &= \frac{x_1}{\lambda_2 - \lambda_1} + \frac{x_2}{\lambda_1 - \lambda_2} \\ \varepsilon_3 &= \frac{\lambda_1 x_1}{\lambda_2 - \lambda_1} + \frac{\lambda_2 x_2}{\lambda_1 - \lambda_2} \end{aligned} \right\} \quad (6)$$

Upon applying the transform (6) to eqns (4), the following canonical form of the equations is obtained:

$$\left. \begin{aligned} \frac{dx_1}{dt} &= \lambda_1 x_1 + u \\ \frac{dx_2}{dt} &= \lambda_2 x_2 + u \\ \frac{dx_3}{dt} &= u \end{aligned} \right\} \quad (7)$$

The case where  $\zeta > 1$

In this case,  $\lambda_1$  and  $\lambda_2$  are real, both negative and distinct. In order to seek the optimum manipulated variable  $u$ , consider the Hamiltonian function<sup>1</sup>:

$$H = (\lambda_1 x_1 + u)\psi_1 + (\lambda_2 x_2 + u)\psi_2 + u\psi_3 \quad (8)$$

where  $\psi_i$ 's ( $i = 1, 2, 3$ ) are defined by

$$\left. \begin{aligned} \frac{d\psi_1}{dt} &= -\frac{\partial H}{\partial x_1} = -\lambda_1 \psi_1 \\ \frac{d\psi_2}{dt} &= -\frac{\partial H}{\partial x_2} = -\lambda_2 \psi_2 \\ \frac{d\psi_3}{dt} &= -\frac{\partial H}{\partial x_3} = 0 \end{aligned} \right\} \quad (9)$$

Following Pontryagin's maximum principle, the optimum manipulated variable  $u$  should maximize the Hamiltonian function of eqn (8) in which the other variables  $\psi_i$  and  $x_i$  ( $i = 1, 2$ ) are considered to be constant. Hence, the optimum manipulated variable  $u$  subject to the constraint eqn (2) is given by

$$u = \text{sign}(\psi_1 + \psi_2 + \psi_3) = \text{sign}(\psi_{10} e^{-\lambda_1 t} + \psi_{20} e^{-\lambda_2 t} + \psi_{30}) \quad (10)$$

where  $\psi_{10}$ ,  $\psi_{20}$  and  $\psi_{30}$  are initial values of the variables  $\psi_1$ ,  $\psi_2$  and  $\psi_3$  at  $t = 0$ . It is clear from eqn (10) that the optimum manipulated variable  $u$  changes its sign twice at most.

Take the desired terminal conditions  $x_1 = x_2 = x_3 = 0$  as the initial conditions, and try to fill up the three-dimensional phase space with the optimum trajectories starting from the origin with decreasing time. Since  $u = \pm 1$ , the following equations are obtained from eqns (7):

$$\left. \begin{aligned} \frac{dx_1}{dx_3} &= \delta \lambda_1 x_1 + 1 \\ \frac{dx_2}{dx_3} &= \delta \lambda_2 x_2 + 1 \\ \frac{dx_3}{dt} &= \delta \end{aligned} \right\} \quad (11)$$

where  $\delta = \pm 1$ . Integration of eqns (11) gives

$$\left. \begin{aligned} \delta \lambda_1 x_1 + 1 &= C_1 e^{\delta \lambda_1 x_3} \\ \delta \lambda_2 x_2 + 1 &= C_2 e^{\delta \lambda_2 x_3} \\ \Delta(\delta x_3) &= \Delta t \end{aligned} \right\} \quad (12)$$

where  $C_1$  and  $C_2$  are constants of integration, and  $\Delta t$  represents an increment of the time. From eqns (12) the trajectory extending from the origin may be expressed by

$$\left. \begin{aligned} \delta \lambda_1 x_1 + 1 &= e^{\delta \lambda_1 x_3} \\ \delta \lambda_2 x_2 + 1 &= e^{\delta \lambda_2 x_3} \end{aligned} \right\} \quad (13)$$

There are two trajectories passing through the origin, one for  $u = \delta$ , the other for  $u = -\delta$ . The portions of the trajectories which tend to the origin with increasing time form a switching curve in the phase space. Consider the portion of the switching curve for  $u = \delta$ . Through every point on this portion of the switching curve passes a trajectory for  $u = -\delta$ . A family of these trajectories forms a switching surface which terminates in the switching curve for  $u = \delta$ . The equation of the switching surface can be found, after a little calculation<sup>3</sup>, as follows:

$$\begin{aligned} \{1 + [1 - e^{\delta \lambda_1 x_3} (1 - \delta \lambda_1 x_1)]^{\frac{1}{\lambda_1}}\}^{\lambda_2} \\ = \{1 + [1 - e^{\delta \lambda_2 x_3} (1 - \delta \lambda_2 x_2)]^{\frac{1}{\lambda_2}}\}^{\lambda_1} \end{aligned} \quad (14)$$

A family of the trajectories for  $u = \delta$  reaching this switching surface fills up a half volume of the phase space; in other words, this surface is the termination of the family of trajectories for  $u = \delta$  whose initial conditions fill in the half phase space.

When a step input is given to the system in equilibrium, the initial conditions are

$$\varepsilon_1 = A, \varepsilon_2 = \varepsilon_3 = 0, \text{ or } x_1 = x_2 = 0, x_3 = -A \quad (15)$$

where  $A$  is the magnitude of the step input. A point corresponding to the initial conditions [eqn (15)] is located on the  $x_3$  axis in the phase space. A family of trajectories starting from points on the  $\delta x_3$  axis forms a surface whose equation is given by

$$(1 + \delta \lambda_1 x_1)^{\lambda_2} = (1 + \delta \lambda_2 x_2)^{\lambda_1} \quad (16)$$

The phase portrait of the optimum third-order system with  $2\zeta = 2.5$  is shown in Figure 2. In this figure, the trajectory  $O S_2$  is the switching curve. The surface  $S_1 O S_2 C$  is a family of the trajectories passing through the switching curve  $O S_2$  and forms the switching surface whose equation is given by eqn (14). The surface  $A B O S_1$ , as given by eqn (16), is formed by a family of trajectories starting from points on the  $\delta x_3$  axis. The straight line  $A S_1$  in Figure 2 is a trajectory which starts from an infinitely distant point from the origin on the  $\delta x_3$  axis.

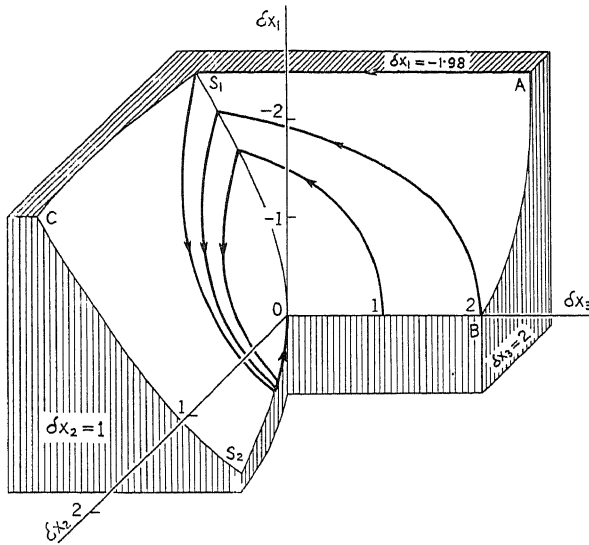


Figure 2. Optimum switching surface and trajectories in the phase space ( $2\zeta = 2.5$ )

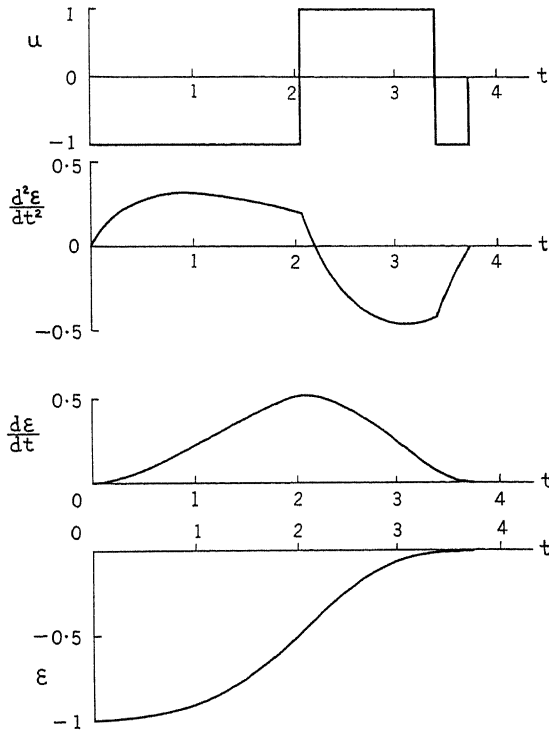


Figure 3. Step response of an optimum third-order servo system ( $2\zeta = 2.5$ )

The third part of eqns (12) enables one to find the total time expended in moving between two points along any combination of trajectories. Figure 3 shows an example of the step response of the system with  $2\zeta = 2.5$ .

The case where  $\zeta = 1$

If  $\zeta = 1$ , then  $\lambda_1 = \lambda_2 = 1$ . Therefore Lurie's transformation [eqn (6)] is inapplicable. In this case use is made of the following transformation:

$$\left. \begin{aligned} \varepsilon_1 &= 2x_1 - x_2 - x_3 \\ \varepsilon_2 &= -x_1 + x_2 \\ \varepsilon_3 &= -x_2 \end{aligned} \right\} \quad (17)$$

Upon applying the transformation (17) to eqn (4), the following set of equations is obtained:

$$\left. \begin{aligned} \frac{dx_1}{dt} &= -x_1 + u \\ \frac{dx_2}{dt} &= -x_1 - x_2 + u \\ \frac{dx_3}{dt} &= u \end{aligned} \right\} \quad (18)$$

Analogously to the preceding case, the optimum manipulated variable  $u$  takes its maximum absolute value under the constraint (2), and changes its sign twice at most. From eqns (18) there is obtained

$$\left. \begin{aligned} \frac{dx_1}{dx_2} &= \frac{\delta x_1 - 1}{\delta x_1 + \delta x_2 - 1} \\ \frac{dx_1}{dx_3} &= -\delta x_1 + 1 \\ \frac{dx_3}{dt} &= \delta \end{aligned} \right\} \quad (19)$$

where  $\delta = \pm 1$ . Each equation of (19) can be integrated separately. Thus

$$\left. \begin{aligned} 1 - \delta x_1 &= C_1 e^{-\delta x_3} \\ \delta x_2 &= (\delta x_1 - 1)(C_2 + \log |1 - \delta x_1|) \\ \Delta(\delta x_3) &= \Delta t \end{aligned} \right\} \quad (20)$$

where  $C_1$  and  $C_2$  are constants of integration. From eqns (20) the trajectory passing through the origin, i.e., the switching curve, may be expressed by

$$\left. \begin{aligned} 1 - \delta x_1 &= e^{-\delta x_3} \\ \delta x_2 &= (\delta x_1 - 1) \log |1 - \delta x_1| \end{aligned} \right\} \quad (21)$$

Proceeding in the same manner as before, the equation of the switching surface is obtained, i.e.,

$$\frac{2[1 - (1 + \delta x_1)e^{-\delta x_3}]^{\frac{1}{2}}}{1 - [1 - (1 + \delta x_1)e^{-\delta x_3}]^{\frac{1}{2}}} \log \{1 + [1 - (1 + \delta x_1)e^{-\delta x_3}]^{\frac{1}{2}}\}$$

$$+ \frac{\delta x_2}{1 + \delta x_1} - \delta x_3 = 0 \quad (22)$$

The step response of the variables may also be obtained as in the preceding case.

The case where  $0 < \zeta < 1$

In this case,  $\lambda_1$  and  $\lambda_2$  are complex conjugate, i.e.

$$\lambda_1 = -\zeta + j(1 - \zeta^2)^{\frac{1}{2}}, \lambda_2 = -\zeta - j(1 - \zeta^2)^{\frac{1}{2}} \quad (23)$$

Lurie's transformation (6) is again applicable. The variables  $x_1$  and  $x_2$  are also complex conjugate and can be written as

$$x_1 = y_1 + jy_2, x_2 = y_1 - jy_2 \quad (24)$$

Lurie's transformation (6) in real form is given by

$$\left. \begin{aligned} \varepsilon_1 &= y_1 + \frac{\zeta}{\omega} y_2 - y_3 \\ \varepsilon_2 &= -\frac{1}{\omega} y_2 \\ \varepsilon_3 &= -y_1 + \frac{\zeta}{\omega} y_2 \\ \omega &= (1 - \zeta^2)^{\frac{1}{2}} \end{aligned} \right\} \quad (25)$$

where

Applying the transformation (25) to eqns (4) yields

$$\left. \begin{aligned} \frac{dy_1}{dt} &= -\zeta y_1 - \omega y_2 + u \\ \frac{dy_2}{dt} &= \omega y_1 - \zeta y_2 \\ \frac{dy_3}{dt} &= u \end{aligned} \right\} \quad (26)$$

The Hamiltonian function for the system of eqns (26) is given by

$$H = (-\zeta y_1 - \omega y_2 + u) \psi_1 + (\omega y_1 - \zeta y_2) \psi_2 + u \psi_3 \quad (27)$$

where  $\psi_i$ 's ( $i = 1, 2, 3$ ) satisfy the following equations:

$$\left. \begin{aligned} \frac{d\psi_1}{dt} &= -\frac{\partial H}{\partial y_1} = \zeta \psi_1 - \omega \psi_2 \\ \frac{d\psi_2}{dt} &= -\frac{\partial H}{\partial y_2} = \omega \psi_1 + \zeta \psi_2 \\ \frac{d\psi_3}{dt} &= -\frac{\partial H}{\partial y_3} = 0 \end{aligned} \right\} \quad (28)$$

Integration of eqns (28) gives

$$\left. \begin{aligned} \psi_1 &= e^{\zeta t} (\psi_{10} \cos \omega t - \psi_{20} \sin \omega t) \\ \psi_2 &= e^{\zeta t} (\psi_{10} \sin \omega t + \psi_{20} \cos \omega t) \\ \psi_3 &= \psi_{30} \end{aligned} \right\} \quad (29)$$

where  $\psi_{10}$ ,  $\psi_{20}$  and  $\psi_{30}$  are initial values of the variables  $\psi_1$ ,  $\psi_2$ , and  $\psi_3$ . The optimum manipulated variable  $u$  which maximizes the Hamiltonian function (27) under the constraint (2) is given by

$$\left. \begin{aligned} u &= \text{sign} (\psi_1 + \psi_3) \\ &= \text{sign} \left\{ \frac{\psi_{30}}{(\psi_{10}^2 + \psi_{20}^2)^{\frac{1}{2}}} + e^{\zeta t} \cos(\omega t - \beta) \right\} \\ \text{where} \quad \beta &= -\tan^{-1} (\psi_{20}/\psi_{10}) \end{aligned} \right\} \quad (30)$$

Take the desired terminal conditions  $y_1 = y_2 = y_3 = 0$  as the initial conditions, and investigate the optimum trajectories which extend from the origin for negative time. Let  $t = -\tau$  ( $\tau > 0$ ), then from eqn (30) the optimum manipulated variable  $u$  is given by

$$u = \text{sign} \{ e^{-\zeta \tau} \cos(\omega \tau + \beta) - C \} \quad (31)$$

where  $C$  and  $\beta$  are parameters dependent on the initial values  $\psi_{10}$ ,  $\psi_{20}$  and  $\psi_{30}$ . For each set of the parameters, an optimum manipulated variable  $u$  and an optimum trajectory extending from the origin are determined<sup>1</sup>. In Figure 4, the damped sinusoidal curve  $x = e^{-\zeta \tau} \cos \omega \tau$  with  $\zeta = 0.1$  is shown. To obtain a sequence of switching instants for a set of the parameters  $C$  and  $\beta$ , the origin of the time  $\tau$  is shifted to the point  $A$  whose coordinates are given by  $\tau = \beta/\omega$  and  $x = C$ . The intersections  $S_1, S_2, S_3, \dots$  of the damped sinusoidal curve and the straight line  $AB$  give the switching instants. Let the length of the time intervals between the switchings be  $\tau_1, \tau_2, \tau_3, \dots$  as indicated in the diagram. When the parameter  $C$  is fixed, the lengths of the time intervals  $\tau_2, \tau_3, \dots$  are constant even if the other parameter  $\beta$  is varied.

From eqns (26),

$$\frac{dy_2}{dy_1} = \frac{\omega y_1 - \zeta y_2}{-\zeta y_1 - \omega y_2 + u} \quad (32)$$

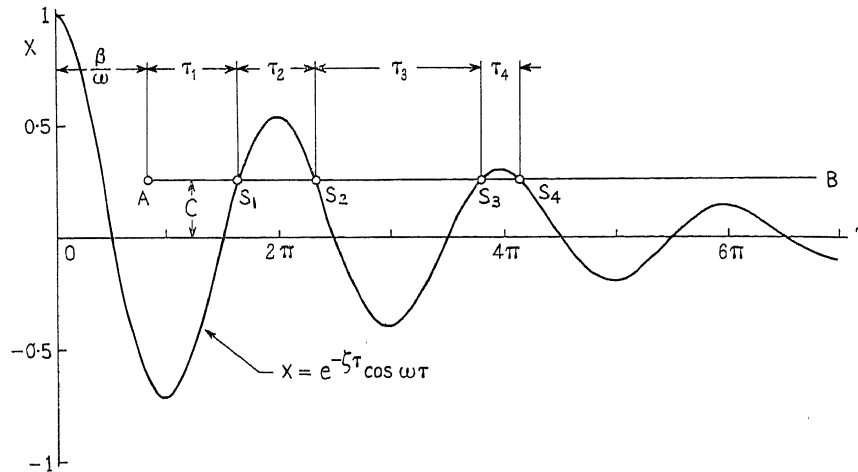


Figure 4. Determination of switching instants ( $\zeta = 0.1$ )

Eqn (32) represents logarithmic spirals around the point  $y_1 = \delta \zeta$ ,  $y_2 = \delta \omega$  ( $\delta = \pm 1$ ) in the  $y_1, y_2$  plane. Figure 5 shows a projection of the optimum trajectories in the  $\delta y_1, \delta y_2$  plane, which are drawn for a certain value of the parameter  $C (= 0.26)$ . The optimum trajectories consist of the arcs of the logarithmic spirals with centre at the point  $M_1$  or  $M_2$ . For this particular value of the parameter  $C$ , the first section of the optimum trajectory extends, for the time interval  $\tau_1$ , from the origin  $O$  to a point on the arc  $ON$ . From this point the second section of the trajectory (with centre at  $M_2$ ) extends for the time interval  $\tau_2$ ,

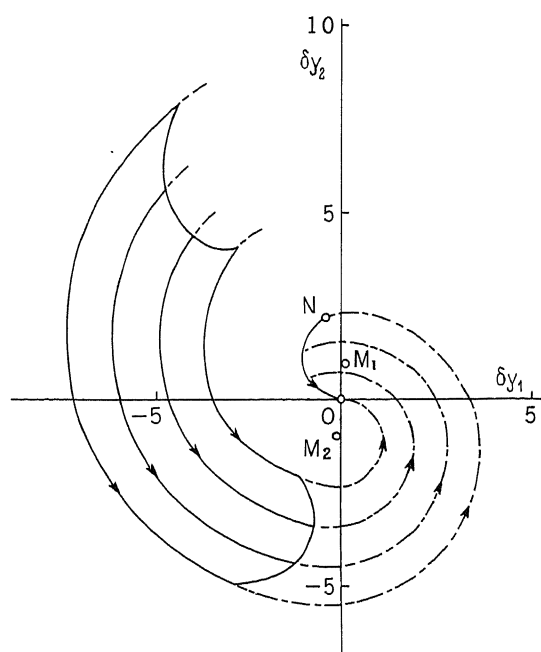


Figure 5. Projection of optimum trajectories in the  $\delta y_1, \delta y_2$  plane ( $\zeta = 0.1$ )

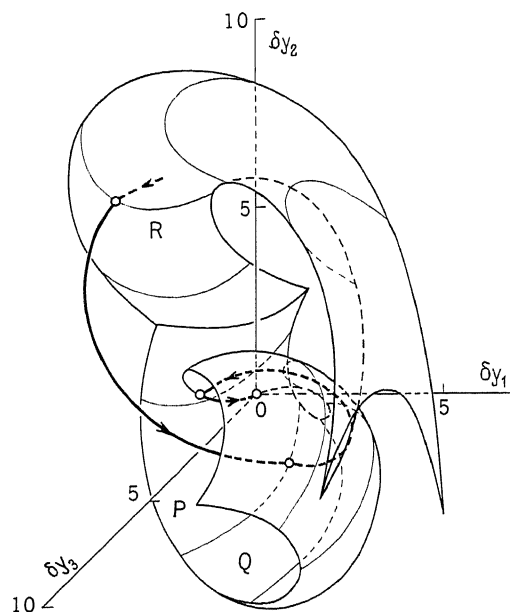


Figure 6. Optimum switching surfaces and a trajectory near the origin ( $\zeta = 0.1$ )

then the third section (with centre at  $M_1$ ) extends for the time interval  $\tau_3$ , and so forth. It is clear from the third equation of (26) that the increment of the variable  $y_3$  is proportional to the angle of rotation of the representative point around the point  $M_1$  or  $M_2$ .

By varying both the parameters  $C$  and  $\beta$ , the phase space may be filled up with optimum trajectories. Figure 6 shows the optimum switching surfaces and an example of the trajectory (shown by thick line) near the origin. In the diagram, the trajectory  $OP$  which extends from the origin is the switching curve, where the switching corresponding to the first intersection

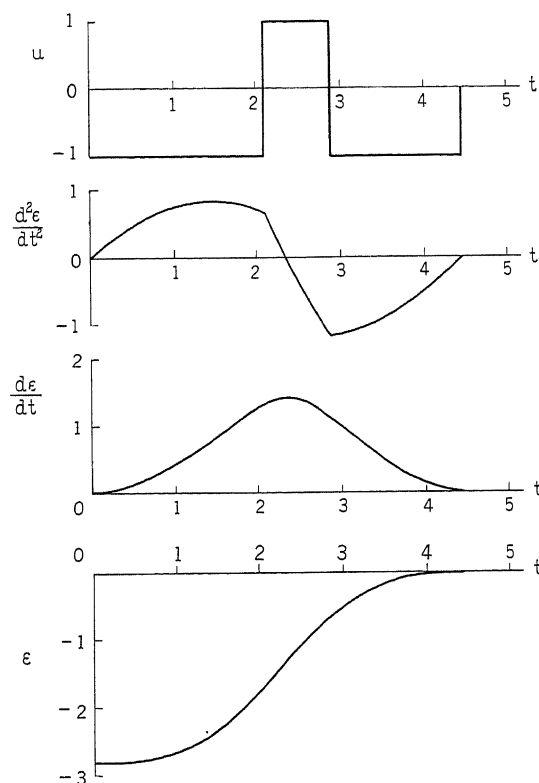


Figure 7. Step response of an optimum third-order servo system ( $\zeta = 0.1$ )

$S_1$  of Figure 4 occurs. The representative point moves along the switching curve  $OP$  during the time interval  $\tau_1$ . A family of trajectories passing through the points of the switching curve  $OP$  constitutes a switching surface  $Q$ , where the switching corresponding to the second intersection  $S_2$  of Figure 4 occurs. The representative point moves on the surface  $Q$  during the time interval  $\tau_2$ . The surface  $R$  is also a switching surface, where the switching corresponding to the third intersection  $S_3$  occurs. The switching surfaces where the switchings corresponding to the intersections  $S_4, S_5, \dots$  occur are omitted in the diagram.

Figure 7 shows an example of the step response of the system with  $\zeta = 0.1$ .

#### Optimum Control of a Non-linear Second-order System

In this section the optimum control of a non-linear system is considered<sup>5</sup>. The system is governed by

$$\frac{d^2 x}{dt^2} + \varepsilon(x^2 - 1) \frac{dx}{dt} + x = u \quad (33)$$

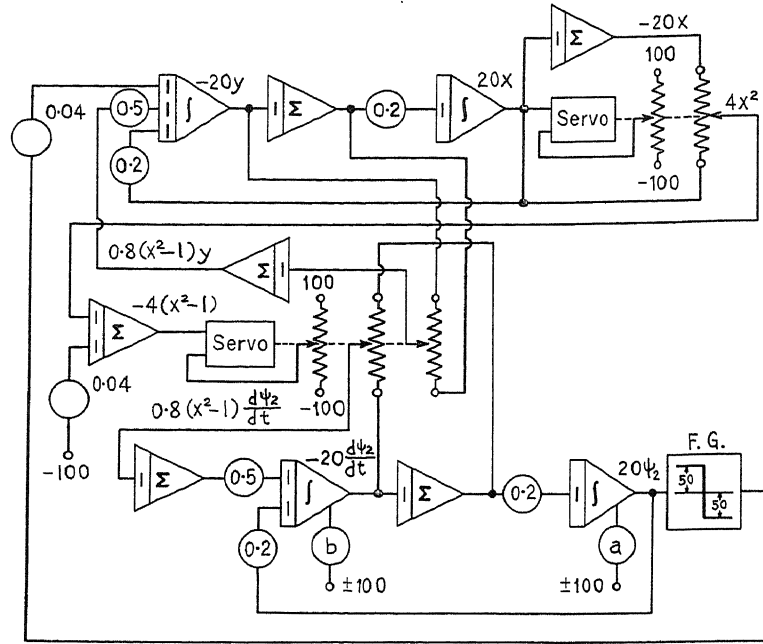


Figure 8. Block diagram of an analogue computer set-up for the solution of eqns (33), (38) and (39) for negative time ( $\varepsilon = 0.1$ )

where  $\varepsilon$  is a positive constant. It is assumed that the manipulated variable  $u$  is subject to the constraint:

$$|u| \leq 0.5 \quad (34)$$

If  $u = 0$ , eqn (33) has a single periodic solution to which all solutions tend, regardless of their initial conditions. In order to counteract this tendency and keep the system at  $x = 0$ ,  $dx/dt = 0$ , the manipulated variable  $u$  has been introduced.  $u$  is to be chosen, subject to the constraint (34), so as to reduce  $x$  and  $dx/dt$  to zero in minimum time.

Eqn (33) is equivalent to

$$\left. \begin{aligned} \frac{dx}{dt} &= y \\ \frac{dy}{dt} &= -x - \varepsilon(x^2 - 1)y + u \end{aligned} \right\} \quad (35)$$

The Hamiltonian function for the system of eqns (35) is given by

$$H = y\psi_1 + \{-x - \varepsilon(x^2 - 1)y + u\}\psi_2 \quad (36)$$

where the variables  $\psi_1$  and  $\psi_2$  are defined by

$$\left. \begin{aligned} \frac{d\psi_1}{dt} &= -\frac{\partial H}{\partial x} = (1 + 2\varepsilon xy)\psi_2 \\ \frac{d\psi_2}{dt} &= -\frac{\partial H}{\partial y} = -\psi_1 + \varepsilon(x^2 - 1)\psi_2 \end{aligned} \right\} \quad (37)$$

The optimum manipulated variable  $u$  which maximizes the Hamiltonian function (36) under the constraint (34) is given by

$$u = 0.5 \operatorname{sign} \psi_2 \quad (38)$$

Eliminating the variable  $\psi_1$  in eqns (37) gives

$$\frac{d^2\psi_2}{dt^2} - \varepsilon(x^2 - 1)\frac{d\psi_2}{dt} + \psi_2 = 0$$

Our purpose is to determine the optimum switching the  $x, y$  phase plane, at which the manipulated variable its sign so as to lead the representative point from a conditions to the origin in minimum time. Since eqns (33) and (39) are hardly solved because of their non-linearity, an analogue computer was made of an analogue computer to obtain the switching curve.

Figure 8 shows a block diagram of an analogue set-up for solving eqns (33), (38) and (39) for negative

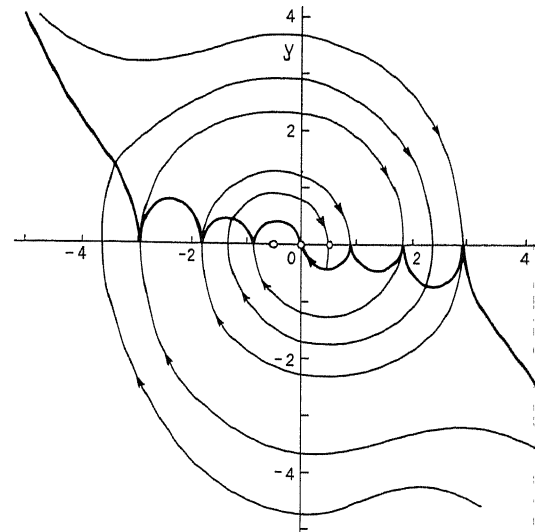


Figure 9. Optimum switching curve and trajectories for a system governed by eqn (33) ( $\varepsilon = 0.1$ )



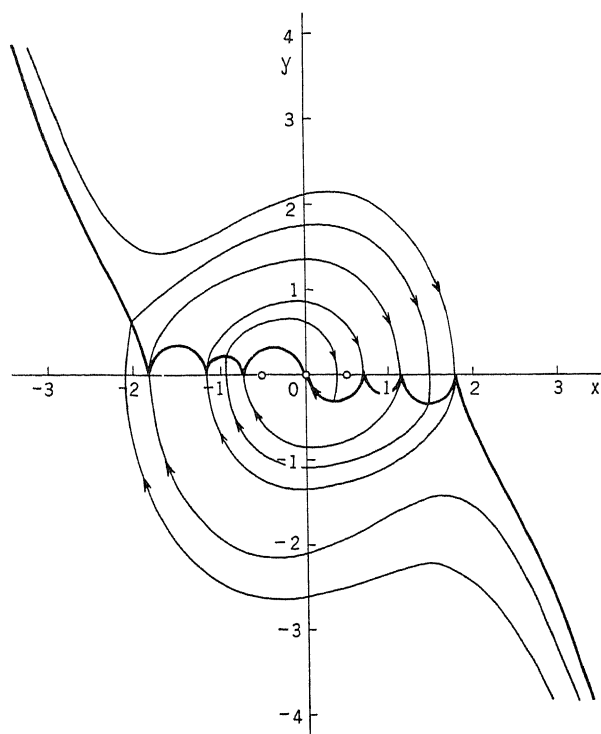


Figure 10. Optimum switching curve and trajectories for a non-linear system governed by eqn (33) ( $\varepsilon = 0.5$ )

Namely, the computer was set up to solve eqns (33), (38) and (39) with  $5t = -\tau$  ( $\tau > 0$ ), where  $\tau$  is computer time and  $t$  is problem time. The desired terminal condition  $x = y = 0$  are taken as the initial conditions for the computer solution. The initial conditions  $\psi_2(0)$  and  $(d\psi_2/dt)_0$  may be varied by the setting of the potentiometers  $a$  and  $b$ . For each set of the initial values  $\psi_2(0)$  and  $(d\psi_2/dt)_0$ , an optimum trajectory starting from the origin is obtained. Figures 9 and 10 show the optimum switching curves and trajectories for  $\varepsilon = 0.1$  and  $\varepsilon = 0.5$ , respectively.

## References

- 1 PONTYAGIN, L. S. Optimal processes of regulation. *Progr. math. Sci. Moscow* 14 (1959) 3. Deutsche Übersetzung 1963. München; R. Oldenbourg
- 2 BOGNER, I., and KAZDA, L. F. An investigation of the optimum switching criteria for higher order contactor servomechanisms. *Trans. Amer. Inst. elect. Engrs* 73 (1954) 118
- 3 HAYASHI, C., and SAKAWA, Y. Optimum switching criteria for higher-order contactor servomechanisms. *J. Inst. elect. Engrs, Japan* 77 (1957) 1601
- 4 LURIE, A. I. *Einige nichtlineare Probleme aus der Theorie der Selbsttätigen Regelung*. 1957. Berlin; Akademie-Verlag
- 5 BELLMAN, R. On the application of the theory of dynamic programming to the study of control processes. *Proc. Symp. on Non-linear Circuit Analysis*. Vol. 6, p. 199, 1956. New York; Polytech. Inst. of Brooklyn

## DISCUSSION

I. FLÜGGE-LOTZ, *Division of Engineering Mechanics, Stanford University, Stanford, California, U.S.A.*

My particular interest in this paper is based on the fact that in 1959 Mih Yin and I investigated time-optimal control of third-order systems

$$\frac{1}{(s+\gamma)(s^2+2\zeta s+1)} \quad \text{with} \quad |\zeta| < |1|$$

We studied solutions in backward time and developed an exact procedure for control in forward time for the case  $\zeta = 0$ ,  $\gamma = 0$  and an approximate procedure for the general case. The control should be realized by digital means because of an iteration for obtaining the initial values of the adjoints. Our report<sup>1</sup> contains many more examples than the printed version<sup>2</sup> published in 1961. In that year H. Titus and I improved (shortened) the iteration procedure<sup>3</sup>.

Did the authors of this paper try to realize the optimum time control (in forward time) of the investigated systems?

## References

- 1 FLÜGGE-LOTZ, I. and YIN, M. On the optimum response of third-order contactor control systems. *Stanford Univ. E.M. Tech. Rep. No. 125 (AFOSR-TN-60-476)* (April 1960)
- 2 FLÜGGE-LOTZ, I. and YIN, M. The optimum response of second-order, velocity-controlled systems with contactor control. *J. bas. Engng, Trans. Amer. Soc. mech. Engrs*, Ser. D, 83 (1961) 59-64
- 3 FLÜGGE-LOTZ, I. and TITUS, H. A. The optimum response of full third-order systems with contactor control. *J. bas. Engng, Trans. Amer. Soc. mech. Engrs*, Ser. D, 84 (1962) 554-558

Y. SAKAWA, *in reply*

Our work partly overlaps that of Professor Flügge-Lotz. We obtained the optimum switching surface of a linear third-order system and the

optimum switching curve of a non-linear second-order system by using backward-time technique; that is, we tried to fill up the phase space with the optimum trajectories extending from the origin with decreasing time. But, now that the optimum switching surface and the optimum switching curve are obtained, we can realize such time-optimal controllers by using non-linear function generators of the analogue or digital type.

For example, the optimum switching curves as shown in Figures 9 and 10 of the paper can be realized by using either a servo-type function generator or a photoformer-type function generator. Furthermore, an optimum switching surface as shown in Figure 2 of the paper can be realized by using a two-variable function generator of the type  $z = f(x, y)$ .

What type of function generator is most effective and proper for the actual controller is another problem; it belongs to the computer problem and is discussed in detail in the fine book of Professor A. A. Feldbaum of the U.S.S.R., 'Computing Devices in Automatic Control Systems'. As for myself, I have had no experience in realizing such controllers yet.

E. PAVLIK, *Siemens and Halske, Karlsruhe, Lassallestr. 9, Germany*

The optimum switching function  $F(e, \dot{e}, \ddot{e})$  is introduced. This function  $F$  is not very good for practical application (noise). Have you made trials of approximation of optimal control with simple functions, for example,  $F(e, \dot{e})$ ?

Y. SAKAWA, *in reply*

In many cases, we can obtain the time derivatives of the error signal from the control system itself. But I agree with Dr. Pavlik that in practice his theory is applicable.

SUN JIAN, *Institute of Mechanics, Academy of Science, Peking, People's Republic of China*

Professor Sakawa has presented an interesting paper but I consider that by his method it is very difficult to solve the synthesis problems with equations of higher order. We have methods which are based upon the conception of isochronous region and we solve such problems, in general, as follows:

Let us consider a linear system with variable coefficients

$$\dot{x} = A(t)x + B(t)u; \quad t \geq 0; \quad u \in U \quad (1)$$

It is well known that if  $U$  is closed and convex, then the isochronous region is bounded, convex and closed set. In my paper, which follows this discussion, it is further proved that the isochronous region is a strictly convex  $n$  dimensional set and continuous in both  $t_0$  and  $T$ .

Furthermore, if for any point  $z$  on boundary  $S$  of  $\Omega$  there is an allowed control  $u(t) \in \dot{U}$  (interior of  $U$ ) satisfying either inequality

$$(-\text{grad } g(z), A(t)z + B(t)u(t)) > 0 \quad (2)$$

or the equality

$$A(t)z + B(t)u(t) \equiv 0 \quad (3)$$

then the isochronous region is strictly monotonously increasing with respect to  $T$ .

By these properties we obtain:

**Theorem 1.** If  $G_{t_0}(\Omega, T)$  is monotonously increasing and there is a  $u(t) \in U$  which transfers the  $x \in \Omega$  into  $\Omega$ , then there exists a unique  $z \in S$  and a smallest  $T$  such that the control

$$u(t) = \text{sgn}(-\text{grad } g(z), \phi(t, t_0 + T)B(t)) \quad (4)$$

is a time optimal control transferring  $x$  into  $\Omega$ , where  $\phi(t, t_0)$  is the inverse fundamental matrix solution of original system. Finally, every  $z$  on  $S$  determines such a time optimal control.

This theorem indicates that Pontryagin's maximum principle is also a sufficient condition for time optimum, if  $\Omega$  and  $U$  satisfy the monotonous condition (2) or (3). However, the inverse conclusion is not true.

Based on the above-mentioned theorem, one can easily find the switching hypersurfaces in phase space as a finite parametrical repre-

sentation. Various examples show that this theorem is useful for synthesizing optimal control.

If we let  $T(x, t)$  denote the optimal response time from  $x$  at instant  $t$  into  $\Omega$ , then  $T(x, t)$  is a single-valued function defined in controllable region of phase space, and we have:

**Theorem 2.** If  $\Omega$  and  $U$  satisfy monotonous condition (2) or (3), then optimal response time  $T(x, t)$  is continuous, piecewise differentiable and almost everywhere satisfies the following partial differential equation

$$\left( \frac{\partial T}{\partial x}, A(t)x + B(t)u \right) = -1 \quad (5)$$

with the boundary condition

$$T(x, t)|_S = 0 \quad (6)$$

This Cauchy's problem has a unique solution definable by  $\text{grad } g(z)$ . The required optimal control function is

$$U(x, t) = \text{sgn} \left( -\frac{\partial T}{\partial x}, B(t) \right) \quad (7)$$

The method of getting a solution for (5)–(6) has been found. Thus, the synthesis problem for linear systems with variable coefficients is solved completely. When the  $\Omega$  consists of only a single point, for instance the origin, the two theorems are still true with only a little modification.

Both the above theorems and proposed methods may be extended to a synthesis of optimal systems with another integral functional criterion. These methods have been well verified by many examples.

#### Reference

- <sup>1</sup> SUNG JIEN, HAN KING-CHING Theory of analysis and synthesis of linear time optimal control systems. *Shuxue Jinzhen* 5, Nr. 4 (1962)

Y. SAKAWA, *in reply*

I think Dr. Sun Jian's method for synthesizing time-optimal systems is very interesting. This method of approach is different to mine, i.e. he has obtained a partial differential equation which the variable  $T(x, t)$  has to satisfy. Once this partial differential equation is solved, the equation of the switching curve is obtained directly.

However, I am afraid that the solution of the partial differential equation will not always be easy.

# Analysis and Synthesis of Time-optimal Control Systems

SUN JIAN and HANG KING-CHING

## Summary

In this paper some new results concerning analysis and synthesis of time-optimal control systems are given. We assume that the motion of the controlled plant is described by linear ordinary differential equations with variable or constant coefficients, and the final states of the system form a convex region of phase space. Geometric properties of the isochronous region of these systems in phase space and the facts connected with these which have fundamental meaning for the solution of the synthesis problem are discussed. Moreover the facts demonstrated which are formulated in theorems have also their independent interests. At the end of this paper is given the method of synthesis of the optimal control function for constant as well as with variable coefficients, and illustrative examples.

## Sommaire

Dans ce rapport nous avons soumis de nouveaux résultats à l'égard de l'analyse et de la synthèse des systèmes de commande à temps optimal. Nous supposons que le mouvement de l'installations commandée est décrit par des équations différentielles linéaires ordinaires avec des coefficients variables ou constants, les états terminaux du système formant une région convexe de l'espace des phases des propriétés géométriques de la région isochrone de ces systèmes dans l'espace des phases, ainsi que les faits y rapportant qui ont une signification fondamentale pour la solution du problème de synthèse, se discutent ici. En outre, les faits démontrés, formulés dans des théorèmes, ont aussi leur intérêt indépendant. A la fin du rapport, nous présentons la méthode de synthèse de la fonction de commande optimale pour des coefficients constants ainsi que variables, avec des exemples explicatifs.

## Zusammenfassung

Die Arbeit gibt einige neue Resultate für Analyse und Synthese von zeitoptimalen Regelsystemen. Wir nehmen an, daß die Bewegung des Regelobjekts durch gewöhnliche lineare Differentialgleichungen mit variablen oder konstanten Koeffizienten dargestellt werden kann und die Endstadien des Systems einen konvexen Phasenraumbereich bilden. Die geometrischen Eigenschaften des isochronen Bereichs dieser Systeme im Phasenraum und die damit verbundenen Tatsachen, die für die Lösung des Synthesenproblems grundlegend sind, werden eingehend behandelt; in Theoremen formuliert sind diese Tatsachen auch als solche von Interesse. Das Ende der Arbeit gibt die Methode für die Synthese der optimalen Regelfunktion für konstante und variable Koeffizienten mit Anschauungsbeispielen.

Recently, theoretical investigations on optimal control processes have made great progress. The proof and the development of the so-called 'maximum principle' opened a new direction for the study of optimal processes. But up to the present the synthesis theories of optimal system are far from complete. In general, the synthesis theory cannot be directly derived from the known properties of optimal processes, because it has not been able to find the required boundary conditions.

The object of this paper is to report briefly some conclusions and a method of computation of the synthesis theory of the

time-optimal control system which are obtained by a detailed study of the geometrical properties of the isochronous region, under the assumption that the motion of the controlled plant is described by a system of linear ordinary differential equations with variable or constant coefficients. The concept of the isochronous region was for the first time introduced<sup>3</sup> for the analysis of limiting the time-optimal control system. This was also successfully used<sup>6</sup> to synthesize a time-optimal second-order system. The relation between the isochronous region and the optimal control function has been indicated<sup>7</sup>.

In this paper some theorems concerning the isochronous region and the time-optimal control for the linear system with variable coefficients are given in Sections I and II, which form the foundation of the synthesis theories stated in Section III. At the same time these theorems also present some independent interest.

Owing to lack of space in this paper, we have only reported some conclusions, and omitted the detailed proofs.

I. Consider a controlled object, the movement of which is described by a system of differential equations with variable coefficients of the following form:

$$\frac{dx}{dt} = A(t)x + B(t)u, \quad t \geq t_0 \quad (1)$$

where  $u = (u^1, u^2, \dots, u^r)$  is a point of a cube  $U: |u^i| \leq 1$ ,  $i = 1, 2, \dots, r$ , in the  $r$ -dimensional Euclidian space ( $r \leq n$ ).

Let  $\Phi(t_0, t)$  be the solution of the matrix equation

$$\frac{d\Phi}{dt} = A(t)\Phi, \quad \Phi(t_0, t_0) = E$$

Then, if an  $\mathcal{L}$ -measurable vector function  $u(t)$ ,  $t_0 \leq t \leq t_0 + T$  with restrictions  $|u^i(t)| \leq 1$ ,  $i = 1, 2, \dots, r$ , is given, the solution of eqn (1) satisfying the initial condition  $x(t_0) = x_0$  is

$$\begin{aligned} x(t) &= \Phi(t_0, t) \left[ x_0 + \int_{t_0}^t \Phi^{-1}(t_0, \tau) B(\tau) u(\tau) d\tau \right] \\ &= \Phi(t_0, t) x_0 + \int_{t_0}^t \Phi(\tau, t) B(\tau) u(\tau) d\tau, \quad t_0 \leq t \leq t_0 + T \end{aligned}$$

All  $\mathcal{L}$ -measurable vector functions  $u(t)$  defined on the interval  $[t_0, t_0 + T]$  with the restrictions  $|u^i(t)| \leq 1$  will be called admissible control functions, and the set of all admissible control functions will be denoted by  $V(t_0, t_0 + T)$ .

Let  $u(t) \in V(t_0, t_0 + T)$  and  $x(t)$  be a solution of eqn (1) corresponding to  $u(t)$  and satisfying the conditions  $x(t_0) = x_0$  and  $x(t_0 + T) = x_1$ . Then we shall call  $x(t)$  a trajectory corresponding to  $u(t)$  and connecting  $x_0$  and  $x_1$ .

**Definition 1.** Let  $G^+(x_0, T)$  ( $G^-(x_0, T)$ ) denote the set of all possible terminal (starting) points of the trajectories with  $u(t) \in V(t_0, t_0 + T)$  and with the fixed starting (terminal) point

$x_0$  at  $t_0(t_0 + T)$ . This set will be called  $T$ -isochronous region for the starting (terminal) point  $x_0$  at  $t_0(t_0 + T)$ . If the terminal points form a set  $\Omega$  then  $G(\Omega, T)$  will denote the  $T$ -isochronous region which takes a point of  $\Omega$  as the terminal point. Obviously

$$G(\Omega, T) = \bigcup_{x_0 \in \Omega} G^-(x_0, T)$$

and each isochronous region generally depends on the choice of  $t_0$ .

Let  $\Omega$  be a bounded and closed-convex set defined by the inequalities

$$g_i(x^1, x^2, \dots, x^n) = g_i(x) \leq 0, \quad i=1, 2, \dots, p$$

where  $g_i$  is continuous and differentiable, and  $\text{grad } g_i(x) \neq 0$  on the boundary  $S$  of  $\Omega$ . Let the symbol  $g(z) = 0$  denote the aggregate of the relations which determine  $S$ , and let  $\text{grad } g(z)$ ,  $z \in S$  denote the normal vector of the supporting hyperplane to  $S$  at  $z$  not lying on the same side with  $S$ . This vector will be called the externally normal vector of  $S$  at  $z$  and its negative vector the internally normal vector.

If  $x_1$  belongs to  $G^+(x_0, T)$  ( $G^-(x_0, T)$ ), then there is  $u(t) \in V(t_0, t_0 + T)$  satisfying

$$x_1 = \Phi(t_0, t_0 + T) \left[ x_0 + \int_{t_0}^{t_0+T} \Phi^{-1}(t_0, \tau) B(\tau) u(\tau) d\tau \right] \quad (2)$$

$$\left( x_1 = \Phi(t_0 + T, t_0) \left[ x_0 + \int_{t_0}^{t_0+T} \Phi^{-1}(t_0 + T, \tau) B(\tau) u(\tau) d\tau \right] \right) \quad (2a)$$

If  $x_1$  belongs to  $G(\Omega, T)$  then there are  $u(t) \in V(t_0, t_0 + T)$  and  $z \in \Omega$  satisfying

$$x_1 = \Phi(t_0 + T, t_0) \left[ z + \int_{t_0}^{t_0+T} \Phi^{-1}(t_0 + T, \tau) B(\tau) u(\tau) d\tau \right] \quad (3)$$

Let  $S^\pm(x_0, T)$ ,  $S(\Omega, T)$  and  $\dot{G}^\pm(x_0, T)$ ,  $\dot{G}(\Omega, T)$  denote the boundary and interior of the isochronous regions respectively.

We now introduce the notion of continuity of isochronous regions as follows:

**Definition 2.** Let  $T_0 > 0$ . If there exists a  $\delta > 0$  for any  $\varepsilon > 0$ , such that for any  $T_1, T_2 \in [t_0, t_0 + T_0]$  with  $|T_1 - T_2| < \delta$ , we always have

$$\sup_{x_2 \in G^\pm(x_0, T_2)} \inf_{x_1 \in G^\pm(x_0, T_1)} \|x_1 - x_2\| < \varepsilon$$

$$\sup_{x_1 \in G^\pm(x_0, T_1)} \inf_{x_2 \in G^\pm(x_0, T_2)} \|x_1 - x_2\| < \varepsilon$$

the  $G^\pm(x_0, T)$  is said to be continuous in  $T$  on  $[t_0, t_0 + T]$ . Similarly, we can define the continuity of  $G(\Omega, T)$  in  $T$ .

**Definition 3.** Let  $T > 0$  be an arbitrary number,  $\psi_0$  any non-zero vector and  $b_i(t)$ ,  $i = 1, 2, \dots, r$ , the  $i$ th column of  $B(t)$ . The control  $u(t)$  with the function

$$u^i(t) = \text{sgn}(\psi_0, \Phi(t, t_0 + T) b_i(t))$$

$$t_0 \leq t \leq t_0 + T, \quad i=1, 2, \dots, r$$

as its  $i$ th component is called a control determined by the vector  $\psi_0$  on  $[t_0, t_0 + T]$  and denoted by

$$u(t) = \text{sgn}(\psi_0, \Phi(t, t_0 + T) B(t)), \quad t_0 \leq t \leq t_0 + T$$

Under the assumption of normality of the system (1)<sup>2</sup>, each component of  $u(t)$  has only a finite number of discontinuities of

the first kind in a finite time interval. Therefore we may assume that  $u(t)$  is left-continuous. Let  $V_s(t_0, t_0 + T)$  denote the set of these controls, then obviously

$$V_s(t_0, t_0 + T) \subset V(t_0, t_0 + T)$$

**Definition 4.** If for  $T_2 > T_1 \geq 0$

$$G^-(x_0, T_1) \subset \dot{G}^-(x_0, T_2)$$

$$G(\Omega, T_1) \subset \dot{G}(\Omega, T_2)$$

then  $G^-(x_0, T)$  and  $G(\Omega, T)$  are said to be monotone in  $T$ .

II. In this section we shall give the fundamental geometrical properties of isochronous regions, and the existence and uniqueness theorem for optimal control, and a sufficient condition of the time-optimal control based on these properties.

Let the system (1) be normal, then for  $T > 0$ ,  $G^\pm(x_0, T)$  is an  $n$ -dimensional bounded closed strictly convex set which changes continuously with the time. Let  $x_1 \in G^\pm(x_0, T)$  be a point to which the control  $u(t) \in V(t_0, t_0 + T)$  corresponds, then for  $x_1$  to be a point on  $S^\pm(x_0, T)$ , it is necessary and sufficient that  $u(t) \in V_s(t_0, t_0 + T)$ . Then  $u(t) \in V_s(t_0, t_0 + T)$  is uniquely determined by  $x_1$ , the non-zero vector  $\psi_0$  of  $u(t)$  is the externally normal vector of  $S^+(x_0, T)$  at  $x_1$ , and the vector  $-\Phi^*(t_0, t_0 + T)\psi_0$  ( $\Phi^*$  denotes the transposed of  $\Phi$ ) is the externally normal vector of  $S^-(x_0, T)$  at  $x_1$ .

By using the continuity and the strict convexity of  $G^+(x_0, T)$  for  $T$ , and the uniqueness of the control  $u(t) \in V_s(t_0, t_0 + T)$  which transfers  $x_0$  into  $x_1 \in S^+(x, T)$ , we can prove the following.

**Theorem 1.** Let  $\Omega \subset R_n$  be a given bounded and closed-convex set, and  $x_0 \in \Omega$  be a given point. If the systems (1) can transfer  $x_0$  into  $\Omega$  by  $u(t) \in V(t_0, t_0 + T)$ , then there exists one and only one control  $u^*(t) \in V_s(t_0, t_0 + T^*)$ ,  $0 \leq T^* \leq T$  which transfers  $x_0$  into  $\Omega$  by the shortest time  $T^*$ .

For the isochronous region  $G(\Omega, T)$ , we can also prove that it is an  $n$ -dimensional bounded and closed-convex set which changes continuously with the time. Let  $x_1 \in G(\Omega, T)$  be a starting point of a certain trajectory  $x(t)$  satisfying the condition  $x(t_0 + T) = z$ ,  $z \in \Omega$ , and corresponding to control  $u(t) \in V(t_0, t_0 + T)$ , then for  $x_1 \in S(\Omega, T)$  it is necessary and sufficient that  $z \in S$  and  $u(t) \in V_s(t_0, t_0 + T)$ . And then the vector which determines  $u(t) \in V_s(t_0, t_0 + T)$  is some  $-\text{grad } g(z)$ , where  $z$  is uniquely determined by the starting point  $x_1$ , hence  $u(t)$  is uniquely determined by  $x_1$  and the vector  $\Phi^*(t_0, t_0 + T)\text{grad } g(z)$  is the externally normal vector of  $S(\Omega, T)$  at  $x_1$ . Conversely, let  $\psi_0$  be an externally normal vector of  $S(\Omega, T)$  at  $x_1$ , then the vector  $\Phi^{-1*}(t_0, t_0 + T)\psi_0$  is an externally normal vector of  $S$  at  $z$  and  $-\Phi^{-1*}(t_0, t_0 + T)\psi_0$  is a vector which determines  $u(t) \in V_s(t_0, t_0 + T)$ .

Hence, when  $S$  is a smooth surface,  $S(\Omega, T)$  is also a smooth surface.

By the above properties, we can prove the following fact. Let  $u^*(t) \in V_s(t_0, t_0 + T)$  be a time-optimal control which transfers  $x_0 \in \Omega$  into  $\Omega$ ,  $x^*(t)$ ,  $t_0 \leq t \leq t_0 + T$ ,  $x^*(t_0 + T) = z$ , be its corresponding trajectory, and  $z$  be the terminal point, then for any  $\alpha > 0$ ,

$$z + \alpha \frac{dx(t)}{dt} \Big|_{t=t_0+T-\alpha} \in \dot{G}^+(x_0, T)$$

$$z - \alpha \frac{dx(t)}{dt} \Big|_{t=t_0+T-\alpha} \in \dot{\Omega}$$

From this we can prove a necessary condition for time-optimal control which transfers  $x_0 \in \Omega$  into  $\Omega$ .

**Theorem 2.** For a control  $u(t)$  and its corresponding trajectory  $x(t)$  to be a time-optimal control and a time-optimal trajectory which transfer  $x_0 = x(t_0) \in \Omega$  into  $\Omega$ , it is necessary that  $T > 0$ , and a point  $z$  on  $S$  such that

$$\left. \begin{aligned} u(t) &= \text{sgn}(-\text{grad } g(z), \Phi(t, t_0 + T) B(t)) \\ (-\text{grad } g(z), A(t)x(t) + B(t)u(t))|_{t=t_0+T-0} &\geq 0 \\ g(z) &= 0 \end{aligned} \right\} \quad (4)$$

We then have for any  $t, t_0 \leq t \leq t_0 + T$ ,

$$\begin{aligned} &(-\Phi^*(t, t_0 + T) \text{grad } g(z), A(t)x(t) + B(t)u(t)) \\ &= C + \int_{t_0+T}^t (-\Phi^*(\tau, t_0 + T) \text{grad } g(z) \\ &\quad A'(\tau)x(\tau) + B'(\tau)u(\tau)) d\tau \end{aligned}$$

where  $C$  is a non-negative constant.

Nevertheless a control  $u(t) \in V_s(t_0, t_0 + T)$  which satisfies Theorem 2 is not always the time-optimal control from  $x_0$  to  $\Omega$ . But we have a sufficient condition as follows.

**Theorem 3.** If the isochronous region  $G(\Omega, T)$  is monotone in  $T$ , then each control of  $V_s(t_0, t_0 + T)$  is a time-optimal control which transfers any point of  $S(\Omega, T)$  into  $\Omega$ .

But when is the isochronous  $G(\Omega, T)$  monotone? We can prove a sufficient condition as follows: If for each  $z \in S$  and  $-\text{grad } g(z)$  there is a control function  $u(t) \in V(t_0, \infty)$  satisfying that (a)  $u(t) \in U$  does not take values from the vertices of  $U$ ,  $t \geq t_0$ ; (b) one of the following two expressions holds:

$$\begin{aligned} &(-\text{grad } g(z), A(t)z + B(t)u(t)) > 0, \quad t \geq t_0 \\ &A(t)z + B(t)u(t) \equiv 0, \quad t \geq t_0 \end{aligned} \quad (5)$$

then the isochronous region  $G(\Omega, T)$  is monotone in  $T$ .

When  $\Omega$  is formed by an isolated point and is the origin  $O$  we obtain from Theorem 3 that when  $T > 0$  the isochronous regions for any normal linear control system (1) take  $O$  as an inner point and are monotone in  $T$ , hence every control of  $V_s(t_0, t_0 + T)$  is a time-optimal control for the transfer of points from  $S^-(0, T)$  to the origin  $O$ .

III. In this section we shall give some results of the synthesis method of time-optimal control for systems with constant or variable coefficients.

At first consider a normal system

$$\frac{dx}{dt} = Ax + Bu \quad (6)$$

where  $A, B$  are constant matrices. The given bounded and convex set  $\Omega$  satisfies the conditions of monotonicity (5).

Let

$$M = \bigcup_{T \geq 0} G(\Omega, T)$$

then it follows from the boundedness, closure, convexity and the monotonicity of  $G(\Omega, T)$  that  $M$  is an open convex set containing  $\Omega$ . A synthesis of time-optimal control, of course, can be carried out only within  $M$ .

The meaning of synthesis-optimal control is to find a control function  $u(x) = (u^1(x), u^2(x), \dots, u^r(x))$ ,  $x \in M$ , which, expressed in terms of the position of points in  $M$ , is such that the trajectory  $x(t)$  of the system

$$\frac{dx}{dt} = Ax + Bu(x)$$

starting from  $x_0 \in M$  reaches  $\Omega$  in a certain time, and where

$$u(x(t)) = (u^1(x(t)), u^2(x(t)), \dots, u^r(x(t)))$$

is a time-optimal control for transferring  $x_0$  to  $\Omega$ .

Since, for the control system (6), each component  $u^i$  of time-optimal control  $u(t)$  takes only one of the values  $+1$  and  $-1$ , the synthesis problem of time-optimal control is reduced to the following for each component  $u^i$  divide  $M$  into two parts  $M_i^+$  and  $M_i^-$  such that  $u^i(t)$  becomes  $+1$  within  $M_i^+$  and  $-1$  within  $M_i^-$ . Let  $M_i^0$  denote the common boundary of  $M_i^+$  and  $M_i^-$ . It can be proved that if  $S$  is smooth, then  $M_i^+$  and  $M_i^-$  are connected open sets, and  $M_i^0 = \overline{M_i^+} \cap \overline{M_i^-}$ ,  $M = \overline{M_i^+} \cup \overline{M_i^-}$ . We call  $M_i^0$  the switching surface of the system (6) for the  $i$ th component  $u^i$  of  $u(t)$ . Clearly, the points on the trajectory where  $u^i(t)$  changes its sign are the points of  $M_i^0$ . In order to solve the synthesis problem, we have only to find the switching surfaces.

Let  $x \in M$ ,  $x \in \Omega$ , and let  $T$  be the optimal time of response from  $x$  to  $\Omega$ , then the time-optimal control  $u(t)$  from  $x$  to  $\Omega$  is

$$u(t) = \text{sgn}(-\text{grad } g(z), e^{-A(t-T)}B), \quad 0 \leq t \leq T \quad (7)$$

where  $z \in S$ . Clearly, when  $t = 0$ , the corresponding point on the optimal trajectory is  $x$ , and the value of time-optimal control is  $\text{sgn}(-\text{grad } g(z), e^{AT}B)$ .

Hence

$$u(x(0)) = \text{sgn}(-\text{grad } g(z), e^{AT}B) \quad (8)$$

Therefore, when  $t = 0$ ,  $x(0) \in S(\Omega, T)$  and the vector  $e^{A^*T} \text{grad } g(z)$  is the externally normal vector of  $(\Omega, T)$  at  $x(0)$ , we have  $x \in M_i^-$  when the inner product of  $b_i$  and externally normal vector of  $S(\Omega, T)$  at  $x$  is positive;  $x \in M_i^+$  when this product is negative;  $x \in M_i^0$  when this product is zero. Hence we have parametric expressions of the switching surfaces  $M_i^0$  as follows:

$$\left. \begin{aligned} x &= e^{-AT} \left( z + \int_0^T e^{A\tau} B \text{sgn}(\text{grad } g(z), e^{A\tau}B) d\tau \right) \\ (\text{grad } g(z), e^{AT}b_i) &= 0 \\ g(z) &= 0, \quad 0 \leq T < \infty \end{aligned} \right\} \quad (9)$$

Solving the two components of  $z$  from the second and third expressions, and substituting them into the first, we get an expression of  $M_i^0$  with  $n-1$  variables including the parameter  $T$ . When  $n$  is small, this method is very convenient.

If  $n = 2$  the switching curve can be expressed by  $T$ ; when  $S$  is a point this method is also useful.

The following is another method for finding  $u(x)$ .

Let  $x \in M$ , then according to Theorem 1 the optimal time of response  $T$  from  $x$  to  $\Omega$  is uniquely determined by  $x$ , hence it is a function of  $x$ , written  $T = T(x)$ . Then we have  $T(z) = 0$ ,  $z \in \Omega$ . We can prove the following:

**Theorem 4.** Let  $G(\Omega, T)$  be monotone in  $T$ , then  $T = T(x)$  is continuous in  $M$ . If  $S$  is smooth and satisfies (5), and if the set of those  $z$  which do not satisfy the first expression of (5) is only a surface on  $S$  whose dimension does not exceed  $n-2$ , then  $T(x)$  is everywhere continuous and differentiable on  $M$  except at the points of  $M_i^0$ ,  $i = 1, 2, \dots, r$  and those of  $S$ . Let  $m$  be a connected sub-region of  $M$  not containing any point of  $M_i^0$ ,  $i = 1, 2, \dots, r$ , then  $T(x)$  satisfies in  $m$  a linear partial differential equation of the first order

$$(\text{grad } T(x), Ax + Bu_0) = -1 \quad (10)$$

where  $u_0$  is any vertex of  $U$ .

If  $T = T(x)$ ,  $x \in M$  was obtained, then the surface determined by  $T(x) = T_0$ ,  $T_0 \geq 0$ , coincides with  $S(\Omega, T_0)$  and the vector  $\text{grad } T(x)$ ,  $x \in S(\Omega, T_0)$ , is an externally normal vector of  $S(\Omega, T_0)$  at  $x$ , namely  $\text{grad } T(x)$  is parallel to the vector  $e^{A^*T} \text{grad } g(z)$ . Therefore we obtain from (8)

$$u(x) = -\text{sgn}(\text{grad } T(x), B) \quad (11)$$

Thus the synthesis problem was completely solved.

In order to solve  $T(x)$  from (10), we ought to determine first of all the value of  $u_0$ . Let a set of points of  $S$  determined by  $(\text{grad } g(z), b_i) = 0$ ,  $i = 1, 2, \dots, r$ , be  $(n-2)$ -dimensional hyper-surfaces, then  $S$  is divided by them into  $2^r$  parts:  $S_1, S_2, \dots, S_{2^r}$ , and on each  $S_i$  the control function  $u(x)$  is completely determined by (8). These values are denoted by  $u_1, u_2, \dots, u_{2^r}$  hereafter.

Then by the Theorem 4, the monotonicity condition and the assumption made on  $S$  above, the Cauchy problem

$$(\text{grad } T(x), Ax + Bu_i) = -1$$

$$T(x)|_{S_i} = 0$$

has a unique solution. We extend this solution  $T = T(x)$  in the direction of increasing  $t$  along the integral curve of the system of equations

$$\frac{dx}{dt} = -Ax - Bu_i, \quad x_0 \in S_i$$

until  $(\text{grad } T(x), b_j) = 0$  for some  $j$  for the first time. Further, we take the surface  $(\text{grad } T(x), b_j) = 0$  as a new initial condition for solving the equation

$$(\text{grad } T(x), Ax + Bu_j) = -1$$

where  $u_j$  is different from  $u_i$  only by a sign of the  $j$ th component. Since, when  $S$  is smooth, it is impossible that there should exist a segment of any time-optimal trajectory on the surface determined by  $(\text{grad } T(x), b_j) = 0$ , Cauchy's problem of this kind has a unique solution.

Continuing in this manner we can obtain a solution  $T = T(x)$  within all  $M$ . Hence  $u(x)$  is completely determined by (11). Clearly the switching surfaces for  $u^i$  are

$$(\text{grad } T(x), b_i) = 0, \quad i = 1, 2, \dots, r$$

Thus the synthesis problem which is proposed here is solved.

**Example 1.** Let the equation of motion of the controlled

$$\frac{dx}{dt} = y, \quad \frac{dy}{dt} = u; \quad |u| \leq 1$$

plant be and suppose terminal states from a circle origin as its centre, and  $\rho$  its radius, defined by

$$g(x, y) = x^2 + y^2 - \rho^2 \leq 0$$

where  $0 < \rho \leq 1$ , we can easily verify that the monotonicity condition (5) is here fulfilled.

Moreover, since any point in the phase space can be brought to the origin by an admissible control, the domain  $M$  of the whole phase plane are superposed, hence the synthesis problem is carried out within the whole plane.

From (4) we have in the upper half-circumference

$$(\text{grad } g(x, y), b) = 2y > 0$$

hence  $u = -1$ . Substitute in (10) and solve the partial differential equation of first order

$$\frac{\partial T}{\partial x} y - \frac{\partial T}{\partial y} + 1 = 0, \quad T(x, y)|_{x^2+y^2=\rho^2} = 0$$

A solution of the above-described Cauchy problem

$$T(x, y) = y - \sqrt{2} \sqrt{\rho^2 + 1 - 2\left(x + \frac{1}{2}y^2\right)} - 2\left(1 - x\right)$$

The switching curve is determined by an algebraic equation

$$\frac{\partial T}{\partial y} = 0$$

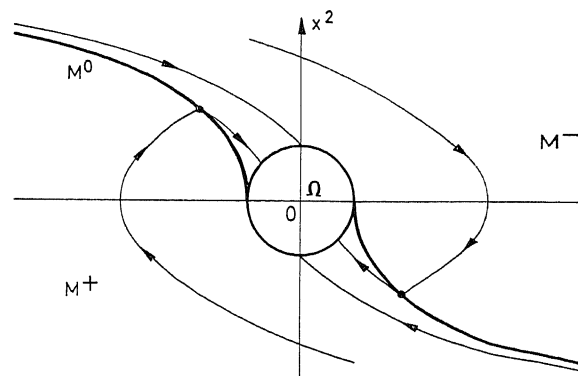


Figure 1

Substituting in (10) the other value of  $u$  and solving the partial differential equation, we can obtain a switching curve on the lower half plane. The phase plane is divided into two parts  $M^+$  and  $M^-$  by two symmetric switching curves  $M^0$  and  $\Omega$ . The time-optimal control function can be written as

$$u(x, y) = \begin{cases} +1, & (x, y) \in M^+ \\ -1, & (x, y) \in M^- \end{cases}$$

The results obtained above can be perfectly generalized to the system with variable coefficients. For the system with parametric expressions (9) of the switching surfaces being

$$\left. \begin{aligned} x &= \Phi(t+T, t)z + \int_t^{t+T} \Phi(\tau, t)B(\tau) \text{sgn}(\text{grad } g(z)) d\tau \\ &\quad \Phi(\tau, t+T)B(\tau) d\tau \\ (\text{grad } g(z), \Phi(t, t+T)b_i(t)) &= 0 \\ g(z) &= 0, \quad 0 \leq T < +\infty \end{aligned} \right\}$$

For the system (1), those switching surfaces depending on  $t$  are denoted by  $M_i^0(t)$ , and the two parts of  $M$  divided by  $M_i^0(t)$  are denoted by  $M_i^+(t)$  and  $M_i^-(t)$ . The necessary and sufficient condition for  $x(t)$  to be a time-optimal trajectory is that  $u^i(t) = +1$ , if  $x(t) \in M_i^+(t)$ ;  $u^i(t) = -1$ , if  $x(t) \in M_i^-(t)$ .

Clearly the synthesis function  $u(x, t)$  is

$$u^i(x, t) = \begin{cases} +1, & x \in M_i^+(t) \\ -1, & x \in M_i^-(t) \end{cases}$$

A partial differential equation of the first order which corresponds to equation (10) is

$$\left( \left( \frac{\partial T}{\partial x^1}, \frac{\partial T}{\partial x^2}, \dots, \frac{\partial T}{\partial x^n} \right), A(t)x + B(t)u_0 \right) + \frac{\partial T}{\partial t} = -1 \quad (10a)$$

where  $T = T(x, t)$  denotes the time of optimal response starting from  $x$  at  $t$  to  $\Omega$ . It is clear that  $T(x, t) \geq 0$  and  $T(x, t)|_S = 0$ . In order to solve  $T(x, t)$  from (10a), we divide first a cylindrical surface  $g(z) = 0$  into  $2^r$  parts in the  $(n+1)$ -dimensional space by the relations

$$(\text{grad } g(z), b_i(t)) = 0, \quad i = 1, 2, \dots, r$$

Then the values of the control functions on these parts are determined for  $t \rightarrow t$ .

Thus we can solve  $T(x, t)$  in a way similar to solving (10). It is obvious that if  $T(x, t)$  was obtained, then

$$u(x, t) = -\text{sgn} \left( \left( \frac{\partial T}{\partial x^1}, \frac{\partial T}{\partial x^2}, \dots, \frac{\partial T}{\partial x^n} \right), B(t) \right)$$

Example 2. We shall find by virtue of (9a) switching surfaces  $M_i^0(t)$ ,  $i = 1, 2$ , of the following system with variable coefficients which takes the origin as the terminal point:

$$\frac{dx^1}{dt} = x^2 + (1 + e^{-t})u^2$$

$$\frac{dx^2}{dt} = -x^1 + (1 + e^{-t})u^1, \quad |u^1| \leq 1, |u^2| \leq 1$$

It is clear that the system is normal. Since the origin is the terminal point, the expressions (9a) of  $M_i^0(t)$  become

$$\left. \begin{aligned} x &= \int_t^{t+T} \Phi(\tau, t) B(\tau) \text{sgn}(\psi, \Phi(\tau, t+T) B(\tau)) d\tau \\ (\psi, \Phi(t, t+T) b_i(t)) &= 0 \\ |\psi| &= 1, \quad 0 \leq T < +\infty \end{aligned} \right\}$$

We solve from the second expression above  $\psi_1 = \pm (\cos T, -\sin T)$ ,  $\psi_2 = \pm (\sin T, \cos T)$ . Owing to the symmetry, we have only to find that part of  $M_1^0(t)$  determined by  $\psi_1 = (\cos T, -\sin T)$ . Since

$$\Phi(t, \tau) = \begin{pmatrix} \cos(\tau-t) & \sin(\tau-t) \\ -\sin(\tau-t) & \cos(\tau-t) \end{pmatrix}$$

the expression of  $M_1^0(t)$  is

$$x = \int_0^T \begin{pmatrix} \cos T & -\sin T \\ \sin T & \cos T \end{pmatrix} \begin{pmatrix} \text{sgn} \cos \tau \\ -\text{sgn} \sin \tau \end{pmatrix} (1 + e^{-t-\tau}) d\tau$$

After the integration is carried out, we have for

$$\frac{n}{2}\pi \leq T \leq \frac{n+1}{2}\pi$$

$x =$

$$\begin{pmatrix} (2n+1)e^{-t} \sum_{i=0}^n e^{-i\frac{\pi}{2}} + \sin \alpha - \cos \alpha - e^{-(t+\frac{n}{2}\pi+\alpha)} \cos \alpha \\ -e^{-t} \sum_{i=1}^n e^{-i\frac{\pi}{2}} + 1 - \cos \alpha - \sin \alpha - e^{-(t+\frac{n}{2}\pi+\alpha)} \sin \alpha \end{pmatrix}$$

where  $\alpha = T - n/2\pi$  (Figure 2).

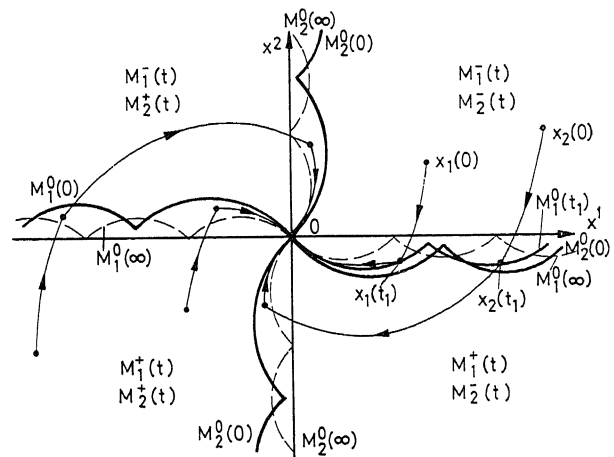


Figure 2

#### References

1. TSIEN, H. S. *Engineering Cybernetics*. 1954
2. PONTRYAGIN, L. S., GAMKRELIDZE, R. V., BOLTYANSKII, V. G. and MISHCHENKO, E. F. *Matematicheskaya Teoriya Optimalnuikh Protssessov*. 1961. Moscow
3. LERNER, A. YA. O predelnom buistrodeistvii sistem avtomaticheskogo upravleniya. *Avtomatika i Telemekhanika* 25, No. 6 (1954)
4. GAMKRELIDZE, R. V. Teriya optimalnuikh po buistrodeistviyn protssessov v lineinuih sistemakh. *Izvest. Akad. Nauk S.S.S.R. (mathematics series)* 22 (1958)
5. FELDBAUM, A. A. O sinteze optimalnuikh sistem s pomoshchyu fazovogo prostvanstva. *Avtomatika i Telemekhanika* 16, No. 2 (1955)
6. BELLMAN, R., GEICHERY, and GROSS. On the bang-bang control process. *Quart. Appl. Math.* 14 (1956) 11-18
7. SUN JIAN. Sintez optimalnuikh sistem na osnovanii polya izokhron. *Izvest Akad. Nauk S.S.S.R., OTN, Energetika i Avtomatika*, No. 5 (1960)
8. SUN JIAN, and HANG KING-CHING. Theory of analysis and synthesis of linear time optimal control systems. *Shuxue Jinzhan* 5, No. 4 (1962)

# Most Recent Development of Dynamic Programming Techniques and Their Application to Optimal Systems Design

R.L. STRATONOVICH

## Summary

Dynamic programming theory will solve, in principle, many of the problems connected with optimal systems synthesis, but the number of complex problems actually solved with this method is not large, because the calculations become more and more difficult the more complicated the problem becomes. Introduction of 'sufficient coordinates', on which the risk function depends, eases the situation, the 'sufficient coordinates' forming the space in which the Bellman equation is considered. In the paper a non-trivial example illustrates the effectiveness of introducing sufficient coordinates. In the example the sufficient coordinates are a combination of *a posteriori* probabilities and the dynamic variable of the controlled plant.

A practical result of the introduction of sufficient coordinates is that the optimal control breaks down in two units, one producing the sufficient coordinates, the other providing optimal control.

## Sommaire

L'emploi de la théorie de la programmation dynamique pourra, en principe, résoudre beaucoup de problèmes de synthèse des systèmes optimaux, mais jusqu'à présent on n'a résolu que peu de problèmes complexes parce que les calculs deviennent de plus en plus difficiles, selon la complexité du problème. L'introduction de 'coordonnées suffisantes', desquelles dépend la fonction de risque, améliore la situation, car celles-ci forment l'espace dans lequel on peut considérer l'équation de Bellman. Dans ce rapport un exemple non courant montre l'utilité d'introduire des coordonnées suffisantes. Celles-ci se composent dans l'exemple d'une combinaison de probabilités *a posteriori* avec la variable dynamique de l'installation réglée.

Un résultat pratique de l'introduction des coordonnées suffisantes, consiste à diviser la commande optimale en deux unités, dont l'une produit les coordonnées suffisantes, et l'autre assure la commande optimale.

## Zusammenfassung

Die Theorie der dynamischen Programmierung kann grundsätzlich viele Probleme der Synthese optimaler Systeme lösen. Sie wurde jedoch bis jetzt noch nicht häufig erfolgreich angewendet, da die Berechnungen mit zunehmendem Komplexitätsgrad immer schwieriger werden. Die Einführung „ausreichender Koordinaten“ von denen die Risikofunktion abhängt, erleichtert die Situation, wobei diese Koordinaten den Bereich angeben, in dem die Bellmansche Gleichung betrachtet wird. Anhand eines schwierigen Beispiels wird die Nützlichkeit der Einführung „ausreichender Koordinaten“ gezeigt. Hierbei sind diese eine Kombination von *a posteriori*-Wahrscheinlichkeiten und der Zeitveränderlichen der Regelstrecke.

Ein praktisches Ergebnis dieses Verfahrens ist die Aufspaltung der Optimal-Regelung in zwei Teileinheiten: eine erzeugt die „ausreichenden Koordinaten“, die andere führt die Optimalregelung durch.

## Introduction: Block-diagram of an Optimal Controller

As this is known<sup>1-4</sup>, dynamic programming theory solves, in principle, a large number of the problems connected with optimal systems synthesis. The applicability of dynamic programming

methods is not impaired by taking into account white Gaussian noise and other random factors in various components—the statistical nature of the signal to be reproduced, imprecise knowledge of it, random influences on the controlled plant, or interference in the feedback circuit (Figure 1). Of course, as the problems grow more complicated, the actual performance of the calculations becomes more and more difficult.

Although the basic principles of dynamic programming were expounded long ago, the number of non-trivial problems of optimal control theory actually solved by this method is not large. This is explained by purely computational difficulties which have to be overcome before a solution is found.

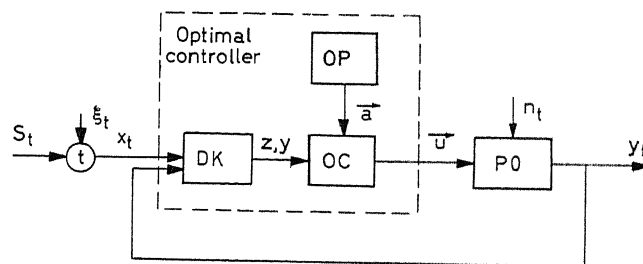


Figure 1. Optimal servosystem. DK: sufficient-coordinates unit; OP: parameter-determination unit; OC: optimal control unit; PO: controlled plant

What has been said confirms the importance of the development of new methods and techniques to increase the effectiveness of the theory and make it easier for concrete results to be obtained.

In complex statistical problems the effective use of the theory becomes possible as a result of the introduction of 'sufficient coordinates' on which the risk function depends. The importance of this concept was noted by Bellman and Kalaba<sup>2</sup>, and the author has clarified and developed it further<sup>5, 6</sup>.

The sufficient coordinates form the space in which the Bellman equation is considered. A non-trivial statistical example is used in this paper to illustrate the effectiveness of the introduction of sufficient coordinates. In the example, the sufficient coordinates are a combination of *a posteriori* probabilities and the dynamic variables of the controlled plant.

In complex statistical problems the introduction of sufficient coordinates has the result that the optimal controller breaks down into at least two consecutive units, each of which is constructed according to its own principles. The first unit DK (Figure 1) produces the sufficient coordinates  $\vec{X}$ . In some dynamic programming problems it is trivial, but in complex statistical problems it may perhaps prove most important. In



the latter, it is synthesized with the aid of methods similar to those of non-linear optimal filtering<sup>7</sup>. In the example considered below, it simply coincides with a unit effecting optimal non-linear filtering.

The signals from the *DK* unit output are sent to a further unit *OC*, which produces the optimal control action. The form of this unit, which converts the sufficient coordinates into a control signal, is found by consideration of the Bellman equation. This unit can be synthesized without great difficulty if the risk function is first found as a solution of the Bellman equation. The most difficult problem is obtain this solution. Therefore, techniques and methods, which make it easier to obtain the solution of this equation, are of interest.

The equation is made far simpler by considering the stationary mode of operation, when the time-dependence and time-derivative are eliminated from the Bellman equation. The corresponding stationary equation was considered by Stratonovich and Shmalgauzen<sup>8</sup>, and the method quoted is also described in this paper. Furthermore, to solve the resulting equation, use is made of the asymptotic step-by-step approximation method, first expounded by the author<sup>9</sup>. This method is convenient for the case of small diffusion terms, and makes it possible to obtain consecutive approximations whose accuracy is determined by the magnitude of the coefficients for the second derivatives in the Bellman equation.

It must be noted that the number of methods for approximate solution of the Bellman equation, which can be thought of for the solution of concrete problems, is practically unlimited; each method is best suited for the solution of problems of a particular type. To these should be added the method for obtaining of a solution on analogue or digital computers. Out of the whole range of methods, a special approximate method will be described and applied to the example under consideration, in the concluding part of the paper. The essence of this method is that the risk function is represented as a function whose appearance is fully determined by a finite number of parameters  $\vec{a}$ . The Bellman equation for the risk function is replaced by a system of equations which specify the evolution of these parameters in inverse time. This system is roughly equivalent to the original Bellman equation.

The unit *OP* (Figure 1) simulates this system of equations and determines the parameters  $\vec{a}$  as a function of time. It operates as a self-contained unit, if measurement of the statistics of the processes and other variables is not carried out in the course of operation, and must finish its work before the start of operation of the main system. If the operating conditions change, then there may be a need for periodic plotting of the process of determination of the parameters by the *OP* unit in application to the new operating conditions. Such a system will belong to the class of adaptive systems. The *OC* unit produces the optimal control action in response to the values of the sufficient coordinates and the risk function parameters corresponding to a given moment of time. The corresponding algorithm is derived from the form of the Bellman equation and the adopted approximation of the risk function.

Usually the transition to a finite number of parameters entails some deterioration of the quality of operation of the system. The greater the number of parameter taken, the higher the accuracy of approximation and the closer the system to optimal, but, on the other hand, the more complicated the *OP* unit. For a

specified number of parameters is important to determine the successful choice of the means of approximation. Here a great deal depends on the ingenuity and inventiveness of the designer. In this paper, one natural means of selecting the parameters is suggested—taken as the parameters are the bottom coefficients of the expansion of the risk function by a suitable full set of functions.

The block diagram of an optimal controller given in the paper is of a basic nature, and in fact not all the units need be there. In some problems the *DK* unit can be left out because of triviality. The *OP* unit can be separated from the system. It can be replaced by a preliminary calculation, and the parameter values can be taken into account once and for all in the synthesis of the *OC* unit. The situation is different if the system itself investigates varying conditions of operation. In that case the signals from the appropriate metering devices must be sent to the units *OP*, *DK* (*OC* if there is no *OP* unit).

### Example—Sufficient Coordinates—Stationary Fluctuation Regime

Let the invariable part of the system—the controlled plant *PO* (Figure 1)—have a transfer function  $K(p)$ . Let the control action  $u$  be limited to the values  $-1 \leq u \leq 1$ . The input signal  $x_t$ , like the output signal  $y_t$ , is assumed to be known accurately. Let the signal on the input  $x_t = s_t + \xi_t$  be the sum of the pulse signal  $s_t = \pm 1$  and disturbance be the normal white noise  $\xi_t$  ( $M\xi_t = 0$ ;  $M\xi_t \xi_{t+\tau} = \kappa \delta(\tau)$ ).

The task of the system is to ensure that the coordinate of the plant  $y_t$  reproduces as accurately as possible the pulse signal  $s_t$ . If  $s_t = 1$ , but  $y_t \neq 1$ , the penalty  $c(1, y_t)$  in a unit of time is taken. The functions  $c(\pm 1, y_t)$  differ. For the step-by-step method, which is used to obtain formula (22), the condition that these functions be differentiable is essential. Henceforward, to make things specific, use will be made of the criterion of the minimum mean square error, which corresponds to the functions

$$c(s, y) = (s - y)^2 \quad (1)$$

It will be assumed that the signal  $s_t$  is *a priori* a symmetrical two-position Markovian process, moreover the *a priori* probabilities  $p_t(\pm 1) = P[s_t = \pm 1]$  satisfy the equations

$$\frac{dp(1)}{dt} = -\frac{dp(-1)}{dt} = -\mu p(1) + \mu p(-1) \quad (2)$$

This means that the pulses and intervals are independent and distributed according to the exponential law  $P[\tau > c] = e^{-\mu c}$ .

It is required to design an optimal controller which produces a control signal  $u_t$  so that the mean penalties are reduced to a minimum. The latter is a function of the sufficient coordinates.

The sufficient coordinates of the given problem will be considered. Their definition, which is given by the author<sup>5, 6</sup> reduces to the requirement of the sufficiency of the selected coordinates in three respects:

(a) Sufficiency for determination of the conditional mean penalties:

$$r_t = M[c_t | x_t, u_t, \tau < t] \quad (3)$$

(b) Sufficiency for indication of the constraints of choice of the control signal and (c) sufficiency for determination of the

future evolution of the sufficient coordinates themselves (for the determination of the probabilities of their future values).

In the given problem the limitations of choice  $|u_t| \leq 1$  at each moment of time  $t$  depend on nothing at all, so point (b) can be disregarded. Point (a) will be considered, and the *a posteriori* probabilities  $w_t(\pm 1) = P[s_t = \pm 1 | x_\tau, \tau < t]$  introduced. Then the mean penalties (3) will be written

$$r_t = c(1, y) w_t(1) + c(-1, y) w_t(-1)$$

Requirement (a) will obviously be satisfied if the sufficient coordinates include the coordinate  $y$  and also the *a posteriori* probability or a magnitude replacing it, say  $z = w(1) - w(-1)$ .

The evolution of the variables of the given problem will be considered. The equation determining the behaviour of  $y_t$  depends on the appearance of the function  $K(p)$ . Obviously

$$\frac{dy_t}{dt} = u_t + n_t \text{ with } K = \frac{1}{p} \quad (4)$$

and

$$\frac{dy_t^2}{dt} + \rho \frac{dy_t}{dt} = \rho u_t + \rho n_t \quad (5)$$

with

$$K(p) = \frac{\rho}{p^2 + \rho p}$$

Assume that  $n_t$  is normal white noise ( $Mn_t = 0$ ;  $Mn_t n_{t+\tau} = N\delta(\tau)$ ). Then in case (4),  $y_t$  will be (with the fixation of  $\{u_\tau\}$ ) a Markovian process, and the probability of the future values  $y_{t+\Delta}$  will be entirely determined by the value at the present moment of time. In case (5), the two-dimensional process  $(y_t, dy_t/dt)$  is Markovian. The probability of the future values is determined by these two magnitudes  $y_t, dy_t/dt$ , and therefore the sufficient coordinates must necessarily include, apart from  $y_t, dy_t/dt$  to satisfy of requirement (c). If, for example, disturbance  $\{n_t\}$  would be a unidimensional Markovian process, then  $n_t$  should be included among the sufficient coordinates.

The mode of variation of  $z_t$  is now found, and it is proved that it does not require the introduction of new sufficient coordinates. The variation of the *a posteriori* probabilities is induced by two causes—*a priori* transfers between states  $s = 1, s = -1$ , and also variation of the *a posteriori* probabilities as a result of supplementary observation of the process  $x_t$ . If there were no observation, the probabilities  $w_t(\pm 1)$  would vary in accordance with eqns (2):

$$\frac{dw_t(1)}{dt} = -\mu w_t(1) + \mu w_t(-1) \quad (6)$$

$$\frac{dw_t(-1)}{dt} = \mu w_t(1) - \mu w_t(-1)$$

If there were no *a priori* transfers, the *a posteriori* probabilities (after the observation  $x_\tau = s + \xi_\tau$  in the interval  $t_0 \leq \tau \leq t$ ) could be expressed through the probabilities  $w_0(\pm 1)$  (before this observation) in accordance with the Beiss formula

$$w_t(s) = \text{const } w[\xi_\tau, t_0 \leq \tau \leq t]_{\xi_\tau = x_\tau - s} \cdot w_0(s) \quad (7)$$

Here  $w[\xi_\tau]$  is the probability distribution for  $\{\xi_\tau, t_0 \leq \tau \leq t\}$  which for white noise, as this is known, has the form

$$w[\xi_\tau] = \text{const exp} \left[ -\frac{1}{2\kappa} \int_{t_0}^t \xi_\tau^2 d\tau \right]$$

Substituting into this  $\xi_\tau = x_\tau - s$ , and relating  $-1/2\kappa \int_{t_0}^t (x_\tau^2 + 1) d\tau$  to the multiplier  $C$ , which does not depend on  $s$ , in accordance with (7), gives

$$w_t(s) = C \exp \left[ \frac{1}{\kappa} \int_{t_0}^t x_\tau s d\tau \right] \cdot w_0(s)$$

From this, differentiation according to  $t$  gives

$$\frac{dw_t(s)}{dt} = \left[ \frac{1}{C} \frac{dC}{dt} + \frac{x_t s}{\kappa} \right] w_t(s) \quad (8)$$

Returning to the case of the *a priori* transfers, eqns (6) and (8) must be combined. This gives

$$\frac{dw(1)}{dt} = -\mu w(1) + \mu w(-1) + \left[ \frac{x_t}{\kappa} + \frac{1}{C} \frac{dC}{dt} \right] w(1)$$

$$\frac{dw(-1)}{dt} = \mu w(1) - \mu w(-1) + \left[ -\frac{x_t}{\kappa} + \frac{1}{C} \frac{dC}{dt} \right] w(-1) \quad (9)$$

The derivative  $1/C \, dC/dt$  is determined from the condition of retention of the norm  $d/dt [w_t(1) + w_t(-1)] = 0$  and proves equal to  $-x_t/\kappa [w(1) - w(-1)]$ . Substituting this value into (9) and transferring to the variable  $z = w(1) - w(-1)$ , gives the equation

$$\frac{dz}{dt} = -2\mu z + \frac{1-z^2}{\kappa} x_t \quad (10)$$

which was derived by the author<sup>7</sup> on the basis of the general theory.

Since in (10)  $x_t = s_t + \xi_t$ , and  $\xi_t$  is white noise, the probabilities of the future values are determined by the value of  $z_t$  and the behaviour of  $s_\tau, \tau > t$ . But since  $s_\tau$  is a Markovian process, its behaviour is determined by the value of  $s_t$ , which is described by the probabilities  $w_t(s_t)$ , that is to say, once again by the coordinate  $z_t$ . Hence the introduction of new variables in accordance with requirement (c) is not necessary.

Equations (4), (5) and (10) make it possible to write an alternative equation or Bellman equation for the given problem. Case (4) will be dealt with first. Introducing the function of minimum future risks

$$S(y, z, t) = \min_{u_\tau, \tau \geq t} M \left\{ \int_t^T c_\tau d\tau \mid y_t, z_t \right\} \quad (11)$$

( $T$  is the time of termination of operation), and compiling the difference of these expressions for the two moments  $t$  and  $t + \Delta$ , gives the equation

$$\frac{\partial S(y, z, t)}{\partial t} + \lim_{\Delta \rightarrow 0} \min_{u_\tau} M \left\{ \frac{S(y_{t+\Delta}, z_{t+\Delta}, t) - S(y_t, z_t, t)}{\Delta} + c_t \mid y_t, z_t \right\} \quad (12)$$

In computing the limit which stands here, a Taylor expansion by the increments  $y_{t+\Delta} - y_t, z_{t+\Delta} - z_t$  will be performed and both the linear and quadratic terms will be taken into account. The differentiability of the risk function is assumed. Eqn (4) gives

$$\lim_{\Delta \rightarrow 0} M \frac{y_{t+\Delta} - y_t}{\Delta} = u_t; \quad \lim_{\Delta \rightarrow 0} M \frac{(y_{t+\Delta} - y_t)^2}{\Delta} = N \quad (13)$$

Computation of the Fokker-Planck coefficients for the second coordinate  $z_t$  is somewhat more complicated. In the process, the equality

$$M\{x_t|z_t\} = M\{s_t|z_t\} = z_t$$

must be taken into account, eqn (10) must be used, and the well-known technique of averaging stochastic eqns (10) must be applied. The result of the averaging has the form

$$\lim_{\Delta \rightarrow 0} M\left\{\frac{z_{t+\Delta} - z_t}{\Delta} \middle| z_t\right\} = -2\mu z_t$$

$$+ \frac{1-z^2}{\kappa} M\{x_t|z_t\} + \frac{\partial}{\partial z_t} \left( \frac{1-z_t^2}{2} \right) \frac{1-z_t^2}{2} = -2\mu z_t \quad (14)$$

Moreover

$$\lim_{\Delta \rightarrow 0} M\frac{1}{\Delta} (y_{t+\Delta} - y_t)(y_{t+\Delta} - y_t) = 0$$

$$\lim_{\Delta \rightarrow 0} M\left\{\frac{(z_{t+\Delta} - z_t)^2}{\Delta} \middle| z_t\right\} = \frac{(1-z_t^2)^2}{\kappa^2} \lim_{\Delta \rightarrow 0} M\frac{(x_{t+\Delta} - x_t)^2}{\Delta}$$

$$= \frac{(1-z_t^2)^2}{\kappa} \quad (15)$$

Hence, eqn (12) adopts the form

$$\frac{\partial S}{\partial t} + \min \left[ \pm \frac{\partial S}{\partial y} \right] - 2\mu z \frac{\partial S}{\partial z} + \frac{N}{2} \frac{\partial^2 S}{\partial y^2} + \frac{(1-z^2)^2}{2\kappa} \frac{\partial^2 S}{\partial z^2}$$

$$+ C(1, y) \frac{1+z}{2} + C(-1, y) \frac{1-z}{2} = 0 \quad (16)$$

The second term can also be written in the form  $-|\partial S/\partial y|$ . To the resulting eqn (16) must be added the boundary conditions. In view of the fact that  $|s| \leq 1$ , only the domain  $|y| \leq 1$  need be considered. Because (16) contains the diffusion term  $1/2 N \partial^2 S/\partial y^2$  on the boundaries  $y = \pm 1$  there must hold the conditions

$$\frac{\partial S}{\partial y}(\pm 1, z, t) = 0 \quad (17)$$

Since  $0 \leq w(s) \leq 1$ , for the second coordinate one has  $|z| \leq 1$ . On the sides  $z = \pm 1$  of the square the diffusion coefficient for the second diffusion member  $1/2 \kappa (1 - z^2)^2 \partial^2 S/\partial z^2$  vanishes. Therefore, instead of the conditions  $\partial S/\partial z = 0$  on these sides the more trivial conditions

$$\left| \frac{\partial S}{\partial z}(y, \pm 1, t) \right| < \infty \quad (18)$$

are satisfied.

$R_+$  will be used to denote the domain of the space of the sufficient coordinates, where  $\partial S/\partial y > 0$ , and correspondingly  $R_-$  where  $\partial S/\partial y < 0$ . The boundary  $\Gamma$  between  $R_+$  and  $R_-$  will be termed the switching line or separatrix; it is to the finding of this line that the calculation of the OC unit (Figure 1) reduces. On this boundary are satisfied the conditions of continuity of the risk function and its first derivatives  $\partial S/\partial y$ ,  $\partial S/\partial z$ . These conditions are a consequence of the diffusion nature of eqn (16). From the continuity of the derivative  $\partial S/\partial y$  there follows the condition

$$\frac{\partial S}{\partial y} = 0 \text{ on } \Gamma \quad (19)$$

Eqn (16) describes the evolution of the risk function with the inverse evolution of time. The role of the initial condition for this

is played by the fixation of the risks at the moment of completing the operation  $S(y, z, T)$ . If there are no special additional considerations, then  $S(y, z, T)$  can be made equal to zero.

The Bellman equation is also derived in a similar way for more complex functions  $K(p)$ . As in case (5), the velocity  $v = \partial y/\partial t$  must be included among the sufficient coordinates. Then the function  $S(y, v, z, t)$  will satisfy the equation

$$\frac{\partial S}{\partial t} + v \frac{\partial S}{\partial y} - \rho v \frac{\partial S}{\partial v} - \rho \left| \frac{\partial S}{\partial v} \right| - 2\mu z \frac{\partial S}{\partial z} + \frac{\rho_2 N}{2} \frac{\partial^2 S}{\partial v^2}$$

$$+ \frac{(1-z^2)^2}{2\kappa} \frac{\partial^2 S}{\partial z^2} + c(1, y) \frac{1+z}{2} + c(-1, y) \frac{1-z}{2} = 0 \quad (20)$$

An important particular problem among the group of problems connected with optimal systems synthesis is the problem of calculating the optimal stationary mode of operation. In this case the operation completing time  $T$  tends to infinity. Then, irrespective the values of the coordinates at the moment  $t$  a stationary fluctuation mode is established in the system, characterized by some mean penalty  $\gamma$  per unit of time. This means that when  $T$  increases, e.g., by  $\Delta t$ , the risk function increases by  $\gamma \Delta t$ .

If the difference  $S(t) - \gamma(T - t)$  is formed and the limit transfer  $T \rightarrow \infty$  performed, the resulting function will not depend on time. In case (4) this function

$$f(z, y) = \lim_{T \rightarrow \infty} [S(y, z, t) - \gamma(T - t)]$$

as can easily be seen in accordance with (16) satisfies the equation

$$\left| \frac{\partial f}{\partial y} \right| + 2\mu z \frac{\partial f}{\partial z} = \frac{N}{2} \frac{\partial^2 f}{\partial y^2} + \frac{1}{2\kappa} (1 - z^2)^2 \frac{\partial^2 f}{\partial z^2} + y^2 - 2yz + 1 - \gamma \quad (21)$$

[here (1) is used]. Moreover the same conditions (17)–(19) are satisfied on the boundaries as before. The solution of eqn (21) makes it possible to find simultaneously the function  $f(y, z)$ , the switching line  $\Gamma$  and the stationary mean penalty  $\gamma$ . The same holds for eqn (20).

### Solving the Bellman Equation

In view of the difficulty of obtaining a precise solution of the alternative equation, various approximate methods can be developed. Some of them will be illustrated, taking eqns (16) and (21) as an example. Of course the methods—for example, the method of parameters—permit generalization to other more complex cases as, say case (20), but then the complexity of the calculations increases markedly. The results obtained with the aid of (16) are also approximately valid for case (20), when  $\rho \gg 1$ , i.e., when the inertia of the controlled plant plays a small part and can be disregarded.

In this case, the optimal control action depends on the variables  $y, z$ , and equals  $u = 1$  in the domain  $R_+$  (correspondingly,  $u = -1$  in  $R_-$ ). Figure 2 shows the approximate location of these domains, and of the switching line; the mean transfer velocities  $M dy/dt$ ,  $M dz/dt$  are also given. An approximate calculation of the switching line was performed in the stationary case (eqn 21), by the asymptotic step-by-step method developed

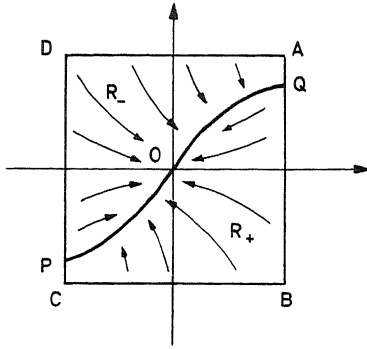


Figure 2. Space of sufficient coordinates.  $POQ$ : separatrix  $\Gamma$ ;  $POQBC$ : domain  $R_+$ ;  $POQAD$ : domain  $R_-$ .

by the author<sup>9</sup>. For the case  $N = 0$ ,  $2\mu < 1$ , the switching line of the first approximation was found to be

$$z_r(y) = y + \frac{2\mu y}{\kappa} \frac{(1-y^2)^2}{1-4\mu^2 y^2} \quad (22)$$

The higher approximations have an order of  $(\mu/\kappa)^2$  and higher.

The second approximate method of solution, which has a wider sphere of application, will be dealt with in greater detail. This method is linked with the determination of the parameters of the risk function, to which corresponds the unit  $OP$  in Figure 1, as was stated in the introduction.

One of the ways of introducing the parameters is the expansion of the risk function according to some preselected suitable system of functions. For the given example these are the functions of the variables  $y$  and  $z$ . Let  $\varphi_0(y), \dots, \varphi_{r-1}(y)$  and  $\psi_0(z), \dots, \psi_{s-1}(z)$  be the selected functions. Then the parameters of the risk function will be the coefficients  $a_{ij}(t)$  of the expansion

$$S(y, z, t) \sim \sum_{i=0}^{r-1} \sum_{j=0}^{s-1} a_{ij}(t) \varphi_i(y) \psi_j(z) \quad (23)$$

Since the above systems of functions are not complete, replacement of the risk function by the expression given usually entails some errors. To make the coefficients  $a_{ij}$  more exact, any criterion is set, e.g., the minimum integral from the square of the difference

$$\int_{-1}^1 \int_{-1}^1 \left[ S - \sum_{ij} a_{ij} \varphi_i \psi_j \right]^2 dy dz = \min$$

will be required.

The variation of this expression leads to a system of linear equations

$$\sum_{i,j} a_{ij} (\varphi_i, \varphi_e) (\psi_j, \psi_m) = (S, \varphi_e \psi_m) \quad (24)$$

$$e=0, \dots, r-1; m=0, \dots, s-1$$

which permits  $a_{ij}$  to be calculated, if  $S(y, z, t)$  is known.

Here is written

$$(\varphi_i, \varphi_e) = \frac{1}{2} \int_{-1}^1 \varphi_i \varphi_e dy;$$

$$(S, \varphi_e \psi_m) = \frac{1}{4} \int_{-1}^1 \int_{-1}^1 S \varphi_e \psi_m dy dz$$

With the aid of the inverse matrices

$$\|c_{ie}\| = \|(\varphi_i, \varphi_e)\|^{-1}; \|c'_{jm}\| = \|(\psi_j, \psi_m)\|^{-1} \quad (25)$$

the solution of system (24) can be written as

$$a_{ij} = \sum_{e,m} c_{ie} c'_{jm} (S, \varphi_e \psi_m) \quad (26)$$

How the equation for the parameters is obtained from the alternative equation will now be shown. Let the latter have the form

$$\frac{\partial S}{\partial t} = \mathcal{F}[S] \quad (27)$$

Differentiating (26) according to time, and substituting (27) into the right-hand side gives

$$\frac{da_{ij}}{dt} = \sum_{e,m} c_{ie} c'_{jm} (\mathcal{F}[S], \varphi_e \psi_m)$$

If the replacement of (23) is performed here, this will give a closed system of equations for the parameters

$$\frac{da_{ij}}{dt} = \sum_{e,m} c_{ie} c'_{jm} (\mathcal{F}[\sum_{p,q} a_{pq} \varphi_p \psi_q], \varphi_e \psi_m) \quad (28)$$

The example being considered will be utilized to illustrate the application of this method. Because of the boundary condition (17) it is convenient to select the functions  $\varphi_i(y)$ , each of which possesses this property  $d\varphi_i/dy (\pm 1) = 0$ . For the second coordinate  $z$ , there is no such condition, so

$$r=s=3; \varphi_0(y) = \psi_0(y) = 1; \varphi_1(y) = \sqrt{2} \sin \frac{\pi y}{2};$$

$$\varphi_2(y) = \sqrt{2} \cos \pi y; \psi_1(z) = z; \psi_2(z) = z^2$$

can be written.

In the given case  $(\varphi_i, \varphi_e) = c_{ie} = \delta_{ie}$ ;

$$\|(\psi_j, \psi_m)\| = \begin{vmatrix} 1 & 0 & \frac{1}{3} \\ 0 & \frac{1}{3} & 0 \\ \frac{1}{3} & 0 & \frac{1}{5} \end{vmatrix}; \|c'_{jm}\| = \begin{vmatrix} \frac{9}{4} & 0 & -\frac{15}{4} \\ 0 & 3 & 0 \\ -\frac{15}{4} & 0 & \frac{45}{4} \end{vmatrix} \quad (29)$$

Since the risk function is symmetrical  $S(y, z, t) = S(-y, -z, t)$  (with symmetrical penalties  $S(y, z, T)$ ), then in expansion (23) there should be present only symmetrical terms

$$S(y, z, t) \sim a_{00} + a_{02} z^2 + a_{11} z \sqrt{2} \sin \frac{\pi}{2} y$$

$$+ (a_{20} + a_{22} z^2) \sqrt{2} \cos \pi y \quad (30)$$

Moreover, putting  $a_{10} = \alpha a_{11}$ ;  $a_{22} = \beta a_{11}$ , it is expedient to make the substitution

$$\left| \frac{\partial S}{\partial y} \right| = \frac{\pi}{\sqrt{2}} |a_{11}| \left| z \cos \frac{\pi}{2} y - 2(\alpha + \beta z^2) \sin \pi y \right|$$

$$\sim \frac{\pi}{\sqrt{2}} |a_{11}| \sum_{ij} \rho_{ij}(\alpha, \beta) \varphi_i(y) z^j \quad (31)$$

where

$$\rho_{ij}(\alpha, \beta) = \sum_{e,m} c_{ie} c'_{jm} \sigma_{em} \quad (32)$$

$$\sigma_{em} = \frac{1}{4} \int_{-1}^1 \int_{-1}^1 \left| z \cos \frac{\pi}{2} y - 2(\alpha + z^2 \beta) \sin \pi y \right| \varphi_e(y) z^m dy dz$$

In addition, within the framework of the selected approximation

$$(1 - z^2)^2 \sim \frac{32}{35} - \frac{8}{7} z^2; y \sim \frac{8}{\pi^2} \sin \frac{\pi}{2} y$$

$$y^2 \sim \frac{1}{3} - \frac{4}{\pi^2} \cos \pi y$$

After the above substitutions, eqn (16), where  $1/2 c(1, y) (1 + z) + 1/2 c(-1, y) (1 - z) = y^2 - 2yz + 1$ , adopts the form

$$\begin{aligned} \sum_{i,j} \frac{da_{ij}}{dt} \varphi_i z^j &= \frac{\pi}{\sqrt{2}} |a_{11}| \sum_{ij} \rho_{ij} \varphi_i z^j \\ &+ 2\mu [2a_{02} z^2 + a_{11} z \varphi_1 + 2a_{22} z^2 \varphi_2] \\ &- \frac{1}{\kappa} \left( \frac{32}{35} - \frac{8}{7} z^2 \right) [a_{02} + a_{22} \varphi_2] \\ &+ \frac{N}{2} \frac{\pi^2}{4} [a_{11} z \varphi_1 + 4(a_{20} + a_{22} z^2) \varphi_2] \\ &- \frac{1}{3} + \frac{2\sqrt{2}}{\pi^2} \varphi_2 + \frac{8\sqrt{2}}{\pi^2} z \varphi_1 - 1 \end{aligned}$$

Equating separately the coefficients of the functions  $\varphi_i z^j$  gives five equations for the  $da_{ij}/dt$  derivatives. The most important of these are the three equations

$$\begin{aligned} \frac{da_{11}}{dt} &= \frac{\pi}{\sqrt{2}} |a_{11}| \rho_{11} \left( \frac{a_{20}}{a_{11}}, \frac{a_{22}}{a_{11}} \right) + 2\mu a_{11} + \frac{\pi^2}{8} N a_{11} + \frac{8\sqrt{2}}{\pi^2} \\ \frac{da_{20}}{dt} &= \frac{\pi}{\sqrt{2}} |a_{11}| \rho_{20} \left( \frac{a_{20}}{a_{11}}, \frac{a_{22}}{a_{11}} \right) - \frac{32}{35} \frac{a_{22}}{\kappa} + \frac{\pi^2}{2} N a_{20} + \frac{2\sqrt{2}}{\pi^2} \\ \frac{da_{22}}{dt} &= \frac{\pi}{\sqrt{2}} |a_{11}| \rho_{22} \left( \frac{a_{20}}{a_{11}}, \frac{a_{22}}{a_{11}} \right) + 4\mu a_{22} + \frac{8}{7} \frac{a_{22}}{\kappa} + \frac{\pi^2}{2} N a_{22} \end{aligned} \quad (33)$$

The switching line is found by equating to zero the derivative (31). The equation of this line has the form

$$4(\alpha + \beta z_r^2) \sin \frac{\pi}{2} y = z_r; z_r = z_r(y) \quad (34)$$

The course of the switching line is determined only by the relations  $\alpha = \frac{a_{20}}{a_{11}}, \beta = \frac{a_{22}}{a_{11}}$  of the parameters entering into (33).

As this is usual in dynamic programming, eqn (33) must be solved for the inverse passage of time. If the inverse time  $t_1 = T - t$  is introduced, the conditions corresponding to the end of operation will look like 'initial' conditions. In the absence of conclusive penalties at the moment  $T$  the corresponding conditions will be null:

$$a_{11} = a_{20} = a_{22} = 0 \text{ when } t_1 = 0 \left( \alpha = \frac{1}{4}, \beta = 0 \right)$$

When a sufficiently long time  $t$  elapses, the mode of operation of the system approaches the stationary. This corresponds to the approach by the parameters  $a_{11}, a_{20}, a_{22}$  of the stationary values  $a_{11}^0, a_{20}^0, a_{22}^0$ . The latter are the solution of the system of three equations obtained by equating to zero expressions (33).

Using (29) and (34), formulae (32) can be brought to the form

$$\begin{aligned} \rho_{11}(\alpha, \beta) &= 3\sigma_{11}; \\ \rho_{10}(\alpha, \beta) &= \frac{9}{4}\sigma_{10} - \frac{15}{4}\sigma_{12}; \\ \rho_{12}(\alpha, \beta) &= \frac{45}{4}\sigma_{12} - \frac{15}{4}\sigma_{10}; \\ \sigma_{ij} &= \frac{1}{2} \int_{-1}^1 \left[ \frac{1 - z_r^{j+2}}{j+2} \cos \frac{\pi}{2} y - 2 \left( \alpha \frac{1 - z_r^{j+1}}{j+1} \right. \right. \\ &\quad \left. \left. + \beta \frac{1 - z_r^{j+3}}{j+3} \right) \sin \pi y \right] \varphi_i(y) dy \end{aligned} \quad (35)$$

For further calculation of the functions  $\rho_{ij}(\alpha, \beta)$  numerical methods can be employed, or use can be made of one or another approximation of the function  $z_r(y)$ .

The solution of the given problem consists in the fact that the unit *OP* (Figure 1) achieves eqns (33) in inverse time, and unit *OC* achieves the switching line (34).

## References

- BELLMAN, R. *Dynamic Programming*. 1960. Moscow; Foreign Languages Publishing House
- BELLMAN, R. and KALABA, R. Dynamic programming and feedback control. *Automatic and Remote Control*. Vol. 1, p. 460. 1961. London; Butterworths
- FELDBAUM, A. A. The theory of dual control, I-IV. *Automat. Telemekh.* 21 (1960), 9, 11; 22 (1961), 1, 2
- STRATONOVICH, R. L. Conditional Markovian processes in problems of mathematical statistics and dynamic programming. *Dokl. Akad. Nauk. SSSR* 140 (1961)
- STRATONOVICH, R. L. On optimal control theory. Sufficient coordinates. *Automat. Telemekh.* 23 (1962), 7
- STRATONOVICH, R. L. Conditional Markovian processes in problems of mathematical statistics, dynamic programming and games theory. *4th All-Union Math. Congr., Leningrad* (1961)
- STRATONOVICH, R. L. Conditional Markovian processes. *Probability Theory and Its Application*, 5 (1960)
- STRATONOVICH, R. L., and SHMALGAUZEN, V. I. Some stationary problems of dynamic programming. *Izv. Akad. Nauk SSSR, Otdel. Tekh. Nauk. Energ. Automat.* 5 (1962)
- STRATONOVICH, R. L. On the optimal control theory. An asymptotic method of solving the diffusion alternative equation. *Automat. Telemekh.* 23 (1962), 11
- STRATONOVICH, R. L. Selected problems of the theory of fluctuations in radio engineering. *Sov. Radio* (1961); Topics in the theory of random noise, Vol. 1, 1963, New York-London; Gordon and Breach

# A Modified Maximum Principle for Optimum Control of a System with Bounded Phase Space Coordinates

S. S. L. CHANG

## Summary

In Pontryagin's maximum principle and in 'bang-bang' control, the manipulated variable is assumed to have no inertia, so that its position changes instantaneously from one value to another. In practice this never happens, for instance in aircraft controls the elevators and ailerons are limited in speed and displacement. The problem is a special case of the more general problem of optimal control in bounded phase space. In this paper a method is given whereby the problem is treated by a limiting process. In place of the rigid bound, a cost function with a multiplied  $K$  is introduced for regions beyond the boundaries in phase space. It is shown that in the limit of  $K$  approaching infinity, both the added cost and the maximum excursion of the optimal path beyond the boundaries approach zero, thus deriving the optimal control condition. The main results of the paper are stated in the form of two theorems. Also considered is the practical significance of the results for systems with multiple saturation.

## Sommaire

Dans le principe du maximum de Pontryagin, ainsi que dans le commande «bang-bang», on suppose que le variable de commande n'a pas d'inertie, de sorte que sa position change instantanément d'une valeur à une autre. Cependant dans la pratique ceci n'arrive jamais: dans les commandes d'avions, par exemple, les gouvernails d'altitude et les ailerons ont des vitesses et des déplacements limités. Ce problème est un cas particulier du problème plus général du commande optimale dans l'espace des phases limité. Dans ce rapport on présente une méthode traitant le problème par un procédé de limitation. Au lieu de la borne rigide, on introduit une fonction de coût avec un  $K$  multiplié pour les régions au delà des limites dans l'espace des phases. On montre qu'à la limite de  $K$  tendant vers l'infini, tant le coût ajouté que le dépassement maximal des limites par le chemin optimal tendent vers zéro, ce qui assure la condition de commande optimale. On établit résultats principaux de ce rapport sous la forme de deux théorèmes. On considère aussi la signification pratique des résultats pour de systèmes à multiple saturation.

## Zusammenfassung

Bei Pontryagin's Maximumprinzip und bei der Zweipunkt-Regelung wird vorausgesetzt, daß die Stellgröße keine Trägheit besitzt, so daß ihre Einstellung augenblicklich wechselt. In der Praxis kommt dies nie vor, so sind z. B. bei der Flugregelung Höhen- und Querruder nach Geschwindigkeit und Ausschlag begrenzt. Dieses Problem ist ein Sonderfall des allgemeineren Problems der optimalen Regelung im begrenzten Phasenraum. In diesem Beitrag wird ein Verfahren angegeben, bei dem das Problem durch einen Begrenzungsprozeß behandelt wird. Anstelle einer starren Grenze wird für einen Bereich außerhalb der Grenzen im Phasenraum eine Kostenfunktion mit Vielfachen von  $K$  eingeführt. Es wird gezeigt, daß, wenn die Grenze von  $K$  gegen Unendlich strebt, sowohl die zusätzlichen Kosten als die größten Abweichungen vom Optimalweg jenseits der Grenzen sich Null nähern. Auf diese Art werden die Bedingungen für optimale Regelung abgeleitet. Die wichtigsten Ergebnisse dieses Beitrags werden in Form zweier Theoreme vorgelegt. Es wird auch die praktische Bedeutung dieser Ergebnisse für Systeme mit mehrfacher Sättigung betrachtet.

## Introduction

The paper aims at removing one essential but impractical condition in the present optimum control theory. In both Pontryagin's maximum principle and the better known 'bang-bang' control, the manipulated variable or rudder is assumed to be inertialess<sup>1-5</sup>. Its position can be changed instantly from  $-a$  to  $a$ . Yet this is never true in actual ships and planes.

The problem can be considered as a special case of a more general problem, that of optimal control in bounded phase space. For instance, in controlling an aeroplane, the elevators and ailerons are limited in both speed and displacement. One way to remove the multiple limits on the movements of the controls is to consider the velocities of the elevators and ailerons as controls only, and to regard the displacements as phase coordinates together with the other dynamical variables of the aeroplane. Then the phase coordinates representing the displacements are bounded.

In a previous paper, Gamkrelidze<sup>6</sup> gave a necessary condition for optimal control in bounded phase space. However Gamkrelidze's condition is quite complicated and difficult to apply, and it has not been proved sufficient for any type of system.

A simpler necessary condition is derived in the present paper for a more restricted class of problem: the boundary in phase space is assumed to be returnable, that is, at every point in the boundary there is an allowed control vector which brings the state vector back to the interior of  $X$ . If the system is linear, and the allowed regions of phase variables and controls are convex, the condition is then both necessary and sufficient.

The method used here is also different from that used by the earlier investigators in that the problem is treated by a limiting process: in place of the rigid bound, a cost function with a multiplier  $K$  is introduced for regions beyond the boundaries in phase space. It is shown that in the limit of  $K$  approaching infinity, both the added cost and the maximum excursion of the optimal path beyond the boundaries approach zero, and the condition for optimal control is thus derived.

The problem and assumptions are stated and the main results of the paper are given in the form of two theorems. Some practical significances of the results for systems with multiple saturation are considered and an outline of the proofs of the theorems is given.

## Vector Representation, Magnitude, Distance

Superscripts are used to denote components of a vector, and subscripts are used to denote components of a covariant vector, e.g.,  $\mathbf{x} = (x^1, x^2, \dots, x^n)$ , and  $\boldsymbol{\psi} = (\psi_1, \psi_2, \dots, \psi_n)$ . The vectors are printed in bold type. Sometimes the 0th component is also included in a vector, and this is shown by underlining, e.g.,  $\underline{\mathbf{x}} = (x^0, x^1, x^2, \dots, x^n)$ .

Subscripts on vectors are used to distinguish different vectors of the same kind. For instance  $x_1$  and  $x_2$  are two points in  $x$  space.

The concept of 'magnitude' of a vector or 'distance' between two points does not enter into the problem nor the theorems. However, in proving the theorems, it is desirable to have these concepts so that closure properties and bounds can be defined or calculated. The magnitude of a vector  $x$  is denoted as  $|x|$  and defined by

$$|x|^2 = (x^1)^2 + (x^2)^2 + \dots + (x^n)^2$$

The same is true for covariant vectors. The distance between two points  $x_1$  and  $x_2$  is  $|x_1 - x_2|$ . The distance between a point  $x_1$  and a region  $X_1$  is

$$\min_{x \in X_1} |x - x_1|$$

### The Problem

The controlled system is described by a set of first order differential equations:

$$\dot{x}^i = f^i(x, u, t) \quad (i=0, 1, 2, \dots, n) \quad (1)$$

where  $x$  and  $u$  are the two vectors  $(x^1, x^2, \dots, x^n)$ , and  $(u^1, u^2, \dots, u^n)$  respectively, and the functions  $f^i$  are single valued, bounded, differentiable with respect to the  $x$ 's with bounded first partial derivatives, and are continuous in the  $u$ 's on a product region  $X_1 \times U_1 \times T_1$ , where  $X_1$ ,  $U_1$ , and  $T_1$  are regions in  $x$  and  $u$  spaces and an interval of  $t$  respectively. The vector  $x$  represents the dynamical state of the system,  $u$  represents the available controls, and  $x^0$  is the return (or cost) function. With  $x(t_1)$  given, the problem is to find a  $u(t)$  such that

$$u(t) \in U \quad (t_1 \leq t \leq t_2) \quad (2)$$

$$x(t) \in X \quad (t_1 \leq t \leq t_2) \quad (3)$$

and

$$R \equiv x^0(t_2) - x^0(t_1) = \text{maximum (or minimum)} \quad (4)$$

where  $X$  and  $U$  are closed regions in  $X_1$  and  $U_1$  respectively. When both eqns (2) and (3) are satisfied,  $u(t)$  is said to be an allowed control, and  $x(t)$  is said to be an allowed path. Among the allowed pairs, the  $u(t)$  and  $x(t)$  which satisfy eqn (4) are said to be optimal control and optimal path respectively.

If  $t_2$  is given but  $x(t_2)$  is unspecified, the problem is said to be one with a fixed time interval or free end point.

If  $x(t_2)$  is given but  $t_2$  is unspecified, the problem is said to be one with a fixed end point. Sometimes the end point  $\xi$  is not fixed but is a function of  $t_2$ . The problem is reducible to one with a fixed end point by using a new state vector  $x'$ :

$$x'(t) \equiv x(t) - \xi(t) \quad (5)$$

If in a problem with fixed end point,  $f^0(x, u, t) = 1$ , and  $R$  is to be a minimum, the problem is then a minimal time problem.

The controlled system is called a linear system if  $f^i(x, u, t)$  can be expressed as linear functions of  $x$  and  $u$ :

$$f^i(x, u, t) \equiv \sum_{j=1}^{j=n} F_j^i(t) x^j + \sum_{k=1}^{k=n'} B_k^i(t) u^k + c^i(t) \quad (i=0, 1, 2, \dots, n) \quad (6)$$

The region  $X$  is assumed to be bounded by simple surfaces on which the condition of returnability is satisfied. To be more specific

(1) A unique normal exists on every boundary point  $x_b$  of  $X$ .

(2) Let  $\eta(x_b)$  denote a unit vector in the direction of the normal at  $x_b$  pointing outward. The partial derivatives  $\partial \eta_i(x) / \partial x^j$  are uniformly bounded.

(3) Every point  $x$  within a certain distance  $d_1$  from  $X$  is on one and only one of the normals  $\eta(x_b)$ . The normal  $\eta(x_b)$  therefore can also be identified as  $\eta(x)$ , where  $x$  may be any point on  $\eta$  within a distance  $d_1$  from the boundary.

(4) There exists a distance  $d_2$ ,  $d_1 > d_2 > 0$ , and a constant  $a_1 > 0$ , such that at every point  $x$  within distance  $d_2$  from  $X$  there are at least some  $u \in U$  satisfying:

$$\sum_{i=1}^{i=n} \eta_i(x) \cdot f^i(x, u, t) < -a_1, \quad u \in U \quad (7)$$

The closed region extending outward from the boundary of  $X$  up to distance  $d_2$  is denoted as  $X_0$ .

$$(5) \quad X + X_0 \subset X_1$$

The optimal control need not be unique, but is assumed to be finite in number. Thus degenerate systems in which the return function is independent of one or more components of the control vector for a finite interval are not under consideration.

The above are the significant assumptions. There are other conditions which are assumed to ensure the closure and limiting properties. These conditions are usually met by a physical system, and are explicitly stated later together with an outline of the proof.

### Results

The results of this paper can be summarized in two theorems: *Theorem 1*—Let  $\hat{u}(t)$  and  $\hat{x}(t)$  for  $t_1 \leq t \leq t_2$  satisfy eqns (1) to (3). A necessary condition for  $\hat{u}(t)$  and  $\hat{x}(t)$  to be optimal in the sense of eqn (4) is that there exist a finite covariant function  $\psi(t)$ , and a function  $\zeta(t)$  satisfying eqns (8) to (10) in the interval  $t_1 \leq t \leq t_2$ :

$$\psi_i + \sum_{k=0}^{k=n} \psi_k \frac{\partial f^k}{\partial x^i} = \zeta(t) \eta_i(\hat{x}) \quad (i=0, 1, 2, \dots, n) \quad (8)$$

where  $\zeta(t) = 0$  if  $\hat{x}(t)$  is an interior point of  $X$ , and  $\zeta(t) \geq 0$  if  $\hat{x}(t)$  is a boundary point of  $X$ ;  $\eta_i$  is the  $i$ th component of  $\eta$ , and  $\eta_0 = 0$ .

$$\sum_{k=0}^{k=n} \psi_k f^k(\hat{x}, \hat{u}, t) = \max_{u \in U} \sum_{k=0}^{k=n} \psi_k f^k(\hat{x}, u, t) \quad (9)$$

$$\psi_0(t) = \text{constant} > 0 \quad (10)$$

( $\psi_0(t) = \text{constant} < 0$  if eqn (4) is to be a minimum)

Furthermore, the following boundary condition on  $\psi(t)$  is satisfied in a free end point problem:

$$\psi_i(t_2) = 0 \quad (i=1, 2, \dots, n) \quad (11)$$

*Theorem 2*—Let  $\hat{u}(t)$  and  $\hat{x}(t)$  for  $t_1 \leq t \leq t_2$  satisfy eqns (1) to (3) of a linear system, and  $X$  be convex. If the conditions stated in Theorem 1 are satisfied by some  $\psi(t)$  and  $\zeta(t)$ , then  $\hat{u}(t)$  and  $\hat{x}(t)$  are optimal control and path functions in the sense of eqn (4).

Theorem 2 can be strengthened by stating that if  $U$  is strongly convex, or if  $U$  is convex and the controlled system is a 'normal system',  $\hat{u}(t)$  and  $\hat{x}(t)$  are the only optimal control and path pair<sup>8</sup>.

### Application to Systems with Multiple Saturation Limits

The optimal control of a system with multiple saturation limits is used here to illustrate a typical application of the two theorems. The controlled system is described by eqn (1), but instead of eqns (2) and (3)

$$|u^i| \leq a_i \quad (12)$$

$$|\dot{u}^i| \leq b_i \quad (13)$$

The problem is reduced to one of bounded phase coordinates by considering  $\dot{u}$  as the control vector, and  $(x, u)$  as an enlarged state vector of  $n + n' + 1$  dimensions. Let the enlarged covariant vector be denoted by  $(\psi, \phi)$  with  $\psi$  corresponding to  $x$  and  $\phi$  corresponding to  $u$ . Then the new Hamiltonian function is

$$H(x, u, \dot{u}, \psi, \phi, t) = \sum_i \psi_i f^i(x, u, t) + \sum_j \phi_j \dot{u}^j \quad (14)$$

Eqn (8) becomes

$$\dot{\psi}_i + \sum_{k=0}^{k=n} \psi_k \frac{\partial f^k}{\partial x^i} = 0 \quad (15)$$

$$\dot{\phi}_i + \sum_{k=0}^{k=n} \psi_k \frac{\partial f^k}{\partial u^i} = \zeta(t) \eta_i \quad (16)$$

Eqn (9) becomes

$$\sum_j \phi_j \dot{u}^j = \max_{|\dot{u}^j| \leq b_j} \sum_j \phi_j \dot{u}^j \quad (17)$$

In an interval in which  $|u^i| < a_i$  (pang interval), eqns (16) and (17) give

$$\dot{\phi}_i = - \sum_{k=0}^{k=n} \psi_k \frac{\partial f^k}{\partial u^i} \quad (18)$$

$$\hat{u}_i = b_i \text{sign } \phi_i \quad (19)$$

In an interval in which  $|u^i| = a_i$  (bang interval)  $\dot{u}_i = 0$ , and eqn (17) can be satisfied only if

$$\phi_i = 0 \quad (20)$$

Eqn (16) gives

$$\begin{aligned} \sum_{k=0}^{k=n} \psi_k \frac{\partial f^k}{\partial u^i} &\geq 0 & \text{if } u^i = a_i \\ &\leq 0 & \text{if } u^i = -a_i \end{aligned} \quad (21)$$

Inequality eqn (21) shows that the optimum  $u^i$  maximizes the Hamiltonian function at least locally in a bang interval.

Summarizing the above, the optimal control is a pang-bang system: either  $u^i$  or  $\dot{u}^i$  is at an extreme value. There are different pang intervals and bang intervals for different components of  $u$ , and generally these do not coincide. From any point  $x$  in phase space, the optimal path to the terminal point may not be unique. However a set of optimal paths can be selected so that there is one and only one optimal path from each point  $x$ . If the system is autonomous, the optimal set and the associated optimum control rate  $\hat{u}$  are time independent. Therefore  $\hat{u}$  can be ex-

pressed as a function of  $x$ . As each component  $\hat{u}$  can take only three values, 0 and  $\pm b_i$ , there are optimum switching boundaries for each component  $\hat{u}^i$ .

Eqn (15) is a differential equation for determining  $\psi$ . Eqns (18) and (20) allow  $\phi$  to be determined from  $\psi$ , and the optimum switching boundaries are given by eqn (19). A numerical example illustrating these points is given by Chang<sup>8, 10</sup>.

### Proof of the Necessary Condition

The proof of Theorem 1 is contained in an earlier report<sup>9</sup>, and only the assumed conditions and essential steps will be given here. Because of condition (3) a distance function  $v(x)$  can be defined as

$$v(x) = 0, \quad \text{if } x \in X$$

$$v(x) = \sum_{i=1}^{i=n} \eta_i(x) (x^i - x_b^i) \quad \text{if } x \in X_0$$

where  $x_b$  is the boundary point of  $X$  upon which  $\eta_i(x)$  falls.

Let the problem defined earlier be denoted as  $P$ . A constructed problem  $P(K)$  is defined as follows:

In eqn (1), the equation for  $i = 0$  is replaced by

$$\dot{x}^0 = f^0(x, u, t) - K[v(x)]^2 \quad (22)$$

Eqn (3) is replaced by

$$x(t) \in X + X_0, \quad t_1 \leq t \leq t_2 \quad (23)$$

The other conditions remain unchanged.

The following Lemmas are then established:

**Lemma 1**—Given  $x(t_1) \in X$ , and a  $K \geq 0$ , there is at least one optimal path in problem  $P(K)$ .

Given the initial point  $x(t_1)$  in the  $n + 1$  dimensional  $x$  space, let the set of all terminal points  $x(t_2)$ ,  $t_2 \geq t_1$ , by an allowed path be denoted  $\Omega(t_2)$ , and the set of all points  $x$  on the allowed paths be denoted  $\Omega(t \leq t_2)$ . The crucial point in proving Lemma 1 is to show that sets  $\Omega(t \leq t_2)$  are closed. This property can be proved by assuming either (a) or (b) as listed below.

(a) The admissible control function  $u(t)$  is measurable, and satisfies a preselected modulus of measurability: there is a finite  $M(\epsilon, \delta)$  for each  $\epsilon > 0$  and  $\delta > 0$ , and for each  $u(t)$  there is a step function  $g(t)$  of no more than  $M$  steps such that

$$|g(t) - u(t)| < \epsilon$$

on  $T_1$  excepting a subset of total measure (or length) less than  $\delta$ .

The alternative (b) is to make assumptions on  $f(x, u, t)$ :

(b.1) The inverse function  $f^{(-1)}(x, z, t)$  is single valued and continuous in all its variables. Thus  $z = f(x, u, t)$  maps  $U$  into a closed region  $Z(x, t)$  in  $z$  space and  $u = f^{(-1)}(x, z, t)$  maps  $Z$  into  $U$ . There is a continuous one to one correspondence between  $z \in Z$  and  $u \in U$  for all  $x$  and  $t$  on  $X_1 T_1$ .

(b.2) The region  $Z(x, t)$  is convex in  $z$  space for all values of  $x$  and  $t$  on  $X_1 T_1$ .

(b.3) The function  $u(t)$  is measurable,  $t_1 \leq t \leq t_2$ .

**Lemma 2**—Given any distance  $d$ ,  $0 < d < d_2$ , a sufficiently large  $M(d)$  can be found such that if  $K > M(d)$ ,  $v(\hat{x}) < d$  for every point  $\hat{x}$  on an optimal path  $\Gamma_0(K)$  of  $P(K)$ .

Lemma 2 is readily proved from eqns (22) and (4).



**Lemma 3**—For sufficiently large  $K$ , it is possible to find a  $M_3$  independent of  $K$  such that

$$K \int_{t_1}^{t_2} v(\hat{x}) dt < M_3 \quad (24)$$

where  $\hat{x}(t)$  describes an optimal path of  $P(K)$ .

Lemma 3 is proved from conditions (2) and (4).

It follows from Lemma 2 that  $\Gamma_0(K)$  stays in the interior of  $X + X_0$  for sufficiently large  $K$ .  $P(K)$  is thus reduced to an optimal control problem without bounds in phase space. Following Pontryagin<sup>4</sup>, eqns (9), (10) and (11) are proved together with an equation for the covariant function

$$\psi_i + \sum_j \psi_j \frac{\partial f^j(x, u, t)}{\partial x^i} = 2\psi_0 K v(\hat{x}) \eta_i(\hat{x}) \quad (25)$$

$$(i=0, 1, 2, \dots, n)$$

Either under assumption (a), or under assumption (b) with a further condition that  $f(x, u, t)$  is separable, a sequence of  $K$  can be selected such that all the terms and relations expressed in eqns (9), (10), (11) and (25) converge to a limit as  $K \rightarrow \infty$ . Because of Lemma 3, the return function for  $P(K)$  converges to the return function for  $P$ , and eqn (25) converges to eqn (8).

#### Proof that the Necessary Condition is also Sufficient for Linear Systems

Let the set  $\hat{u}(t)$ ,  $\hat{x}(t)$ , and  $\psi(t)$  together satisfy eqns (1) to (3) and (8) to (11) with the linear form of  $f^i(x, u, t)$  given in eqn (6). Let  $u(t)$  and  $x(t)$  be another allowed control and allowed path pair. Then it is shown from eqns (1) and (8):

$$\begin{aligned} & \frac{d}{dt} \sum_i \psi_i(t) [\hat{x}^i(t) - x^i(t)] \\ &= \sum_i \zeta(t) \eta_i(\hat{x}) [\hat{x}^i(t) - x^i(t)] \\ &+ \sum_i \sum_k \psi_i(t) B_k^i(t) [\hat{u}^k(t) - u^k(t)] \end{aligned} \quad (26)$$

Note that for the problem with fixed end point, eqn (11) is replaced by

$$x(t_2) = \hat{x}(t_2) \quad (27)$$

Integrating eqn (26) from  $t_1$  to  $t_2$  and making use of eqns (11) or (27) give

$$\begin{aligned} & \psi_0 [\hat{x}^0(t_2) - x^0(t_2)] \\ &= \int_{t_1}^{t_2} \sum_i \zeta(t) \eta_i(\hat{x}) [\hat{x}^i(t) - x^i(t)] dt \\ &+ \int_{t_1}^{t_2} \sum_i \sum_k \psi_i(t) B_k^i(t) [\hat{u}^k(t) - u^k(t)] dt \end{aligned} \quad (28)$$

It follows from the convexity of  $X$  and eqn (9) that both integrals on the right-hand side of eqn (28) are non-negative. Therefore

$$\psi_0 [\hat{x}^0(t_2) - x^0(t_2)] \geq 0 \quad (29)$$

The optimality of  $\hat{x}(t)$  follows from eqns (10) and (29).

For the problem with fixed end point, the optimal system and the system to be compared generally do not terminate at the same time. In order to integrate to the same upper time limit  $t_2$ , an auxiliary condition is necessary: Let  $\xi$  denote the terminating point. It is assumed that there is a

$$u \in U \text{ such that } f^i(\xi, u, t) = 0 \quad (30)$$

The above theorem can also be generalized to linear systems with non-linear return function.

$$F^0(x, u, t) = -g_1(x, t) + g_2(u, t) \quad (31)$$

where  $g_1(x, t)$  is a convex function of  $x$  for every  $t$ . Eqn (6) then holds only for  $i = 1, 2, \dots, n$ . Eqn (26) becomes

$$\begin{aligned} & \frac{d}{dt} \sum_{i=1}^n \psi_i(t) [\hat{x}^i(t) - x^i(t)] \\ &= \sum_{i=1}^n \left\{ \zeta(t) \eta_i(\hat{x}) [\hat{x}^i(t) - x^i(t)] + \sum_k B_k^i(t) [\hat{u}^k(t) - u^k(t)] \right. \\ &+ \left. \psi_0 \frac{\partial g_1(\hat{x}, t)}{\partial \hat{x}^i} [\hat{x}^i(t) - x^i(t)] \right\} \\ &+ \psi_0 [-g_1(\hat{x}, t) + g_2(\hat{u}, t) + g_1(x, t) - g_2(u, t)] \end{aligned} \quad (32)$$

On the right-hand side of eqn (32)

$$\begin{aligned} & \sum_{i=1}^n \zeta(t) \eta_i(\hat{x}) [\hat{x}^i(t) - x^i(t)] \geq 0 \\ & \sum_{i=1}^n \sum_k \psi_i B_k^i(t) [\hat{u}^k(t) - u^k(t)] + g_2(\hat{u}, t) - g_2(u, t) \geq 0 \\ & \sum_{i=1}^n \psi_0 \frac{\partial g_1(\hat{x}, t)}{\partial \hat{x}^i} [\hat{x}^i(t) - x^i(t)] - \psi_0 [g_1(\hat{x}, t) - g_1(x, t)] \geq 0 \end{aligned}$$

because of the convexity of  $X$ , eqn (9) and the convexity of  $g_1(\hat{x}, t)$  respectively. Integrating eqn (32) from  $t_1$  to  $t_2$  gives the more general version of theorem 2.

The work upon which this paper was based was sponsored by the Office of Scientific Research, Air Research and Development Command, Washington 25, D.C., under Grant No. AF-AFOSR-62-321.

#### References

- 1 McDONALD, D. C. Nonlinear techniques for improving servo performance. *Proc. nat. Electron. Conf.* 6 (1950) 400
- 2 FLÜGGE-LOTZ, I. *Discontinuous Automatic Control*. 1953. Princeton; Princeton University Press
- 3 LA SALLE, J. P. Time Optimal Control Systems. *Contributions to the Theory of Nonlinear Oscillations*, vol. 5, *Ann. Math. Stud.* (1960), and *Proc. nat. Acad. Sci., Wash.*, 45 (1959) 573
- 4 BOLTYANSKII, V. G., GAMKRELIDZE, R. V. and PONTRYAGIN, L. S. The theory of optimal processes I, maximum principle. *Bull. Acad. Sci. U.R.S.S. Ser. Mat.*, 24 (1960). *Transl., Space Techn. Lab.*, Rep. No. 9810. 32-01 (October 1960)
- 5 CHANG, S. S. L. *Synthesis of Optimum Control Systems*. 1961. New York; McGraw-Hill
- 6 GAMKRELIDZE, R. V. Optimal control processes with restricted phase coordinates. *Bull. Acad. Sci. U.R.S.S. Ser. Mat.* 24 (1960) 315
- 7 CHANG, S. S. L. Optimal control in bounded phase space. *AFOSR Rep. No. 1238* (1961)
- 8 CHANG, S. S. L. Optimal control in bounded phase space. *Automatica* 1, No. 1 (1963) January
- 9 CHANG, S. S. L. An extension of Ascoli's theorem and its applications to the theory of optimal control. *AFOSR Report 1973*, January 1962
- 10 CHANG, S. S. L. Minimal time control with multiple saturation limits. *Inst. Radio Engrs, N. Y. International Convention Record*, 1962

## DISCUSSION

W. DE BACKER, *C.C.R. Euratom Cetis, Ispra/Varese, Italy*

I am very interested in the contribution of Professor Chang, not only from the theoretical point of view but especially for the computational aspects, with which we are very much concerned in the sections of analogue and digital computing of CETIS, the Scientific Data Processing Center (EURATOM) in Italy.

In fact, the theorems on his  $P(K)$  problem (about which we think in terms of penalty problem in connection with the ideas of Courant, Moser and Kelley) result in an approximation (for finite  $K$ ), by implicit computing, to the Lagrange multipliers as they were introduced by Gamkrelidze. This implicit computing technique, which is essentially a feedback approach, may be promising not only for the digital computer, but surely for analogue computers, where the constraint violations can be kept very small by the high-gain amplifiers.

I shall be very glad to show Professor Chang the first results we had with this technique when applied to a very simple test problem. Our future efforts will be concerned with the stability problems of this technique and I would appreciate any comments Professor Chang might care to give.

S. S. L. CHANG, *in reply*

I am very interested in the computational aspects as mentioned by Dr. De Backer. In fact it is the essential idea underlying the theoretical approach. I believe that his work, when published, will be an interesting contribution.

The question concerning the stability problem can be interpreted in two ways. I do not know which interpretation is meant, therefore I will answer both.

(1) Stability of the optimum trajectory as the value of  $K$  varies: I have proved the existence of a limit of the optimum trajectories only for a selected sequence of  $K$  as  $K$  increases, but not for any value of  $K$  as  $K$  increases. However, only in special problems made up by mathematicians the former may not imply the latter. In most engineering problems it is quite safe to assume the equivalence of the two. As  $K$  increases, the optimum trajectory converges.

(2) The stability of the hardware which is made according to this principle: If  $u$  is expressed as a function of the initial condition and time, the system may or may not be stable. But if  $u$  is expressed as a function of the state variables, then the system is very definitely stable by Liapunov's theorem. I shall outline the process as follows: for any final point or condition there is one or more optimal trajectories passing through each point in state space. Therefore, there is one or more optimal values of  $u$  defined for each point in state space. We may select any one among the few and obtain  $u$  as a function of  $x$ . Then a Liapunov function can be defined by:

$$V(\hat{x}_a) = \int_{t_a}^{t_2} f^0(\hat{x}(t), \hat{u}(x(t))) dt$$

where  $\hat{x}_a = \hat{x}(t_a)$  and  $\hat{x}(t_2)$  is the desired terminal point.

# Optimum and Quasi-optimum Control of Third- and Fourth-order Systems\*

I. FLÜGGE-LOTZ and H.A. TITUS, JUNR.

## Summary

Pontryagin's maximum principle is used for computing the optimum control function  $u(t)$  for a given plant and a given performance criterion. If  $u(t)$  is bounded, the control is of the bang-bang type in many cases. If  $u(t)$  is expressed as the function of the state variables, that means,  $u(t) = \text{sgn } f(x^i)$ , the equation  $f(x^i) = 0$  determines the switching surface in the state space. In general, these surfaces are not given by simple analytic functions; in particular, not if the transfer function of the plant contains complex poles. If the initial state is considered to be given by a disturbance of the desired state, initial error and error derivatives are finite; in this case the final state is given by error and error derivatives being zero, and the switching surface goes through the origin of the error phase space.

Based on experiences with second-order plants, a systematic attempt has been made to approximate the exact switching surfaces for third-order plants. There is an approximation of the surface portion close to the origin (the so-called 'inner' surface) and an approximation of the larger portion of the switching surface which is not close to the origin (the 'outer' surface). Examples show the use of these surfaces; their results are compared to results with exactly optimum switching. They agree well.

The extension to fourth-order systems is indicated.

## Sommaire

Le principe du maximum de Pontryagin est utilisé pour calculer la fonction de commande optimale  $u(t)$  pour un système donné et pour un critère de performance choisi. Si  $u(t)$  est borné, la commande est bien souvent du type relais. Si  $u(t)$  est exprimé en fonction des variables de l'espace de phase, c'est-à-dire  $u(t) = \text{sgn } f(x^i)$ , l'équation  $f(x^i) = 0$  détermine la surface de commande dans l'espace de phase. En général ce ne sont pas des fonctions analytiques simples qui déterminent ces surfaces; elles sont très compliquées quand le système considéré a des pôles complexes. Si dans l'état final l'erreur et sa dérivée sont égales à zéro, cette surface passe par l'origine de l'espace de phase.

A partir des résultats obtenus pour les systèmes du second ordre les auteurs ont développé une théorie d'approximation des surfaces de commande des systèmes du troisième ordre. Ils donnent deux approximations de la surface de commande, l'une pour la zone autour de l'origine des coordonnées de phase (nommée 'surface intérieure'), l'autre pour la partie de cette surface loin de l'origine (nommée «surface extérieure»). Des exemples d'emploi des surfaces approximatives sont donnés; les performances de commande obtenues par cette méthode sont comparées aux performances obtenues en employant les surfaces de commande exactes. La concordance est bonne.

L'extension aux systèmes du quatrième ordre est proposé.

## Zusammenfassung

Das Pontryaginsche Maximumprinzip wird zur Berechnung der optimalen Regelfunktion  $u(t)$  für ein gegebenes System und ein gegebenes Gütekriterium benutzt. Ist  $u(t)$  beschränkt, so besitzen solche Regelungen in vielen Fällen Zweipunktverhalten. Stellt man  $u(t)$  als

\* This research is supported by the Air Force Office of Scientific Research, Air Research and Development Command under Contract AF 49 (638)-513.

Funktion der Koordinaten des Phasenraumes dar, das heißt,  $u(t) = \text{sgn } f(x^i)$ , so bestimmt die Gleichung  $f(x^i) = 0$  die Schaltfläche im Phasenraum. Im allgemeinen genügen diese Schaltflächen keiner einfachen analytischen Funktion, besonders dann, wenn die Übertragungsfunktion des Systems komplexe Pole aufweist. Für einen Anfangszustand, der sich aus Störungen vom gewünschten Zustand ergibt, sind die Abweichung und die Ableitungen der Anfangswerte endlich; in diesem Fall verschwinden Fehler und seine Ableitungen, und die Schaltfläche geht durch den Koordinatenursprung des Phasenraumes.

Auf Grund der Erfahrungen mit Systemen zweiter Ordnung wurde ein systematischer Versuch unternommen, eine Näherung für die exakten Schaltflächen von Systemen dritter Ordnung zu finden. Es ergab sich eine Näherung in der Umgebung des Koordinatenursprungs («innere» Schaltfläche) und eine andere Näherung für den großen Bereich der Schaltfläche, dessen Punkte weiter vom Koordinatenursprung entfernt sind («äußere» Schaltfläche). Beispiele zeigen die Anwendung dieser angenäherten Schaltflächen. Das Verhalten der Regelungen mit den angenäherten und den exakten Schaltflächen wird verglichen; es ergibt sich eine gute Übereinstimmung.

Auf eine Erweiterung dieser Gedanken auf Systeme vierter Ordnung ist hingewiesen.

With the help of Pontryagin's maximum principle a designer can determine for a given plant a control system which is optimum for a given performance criterion. Many papers have been devoted to the investigation of zeroing initial disturbances in minimum time, because this seemed a desirable performance. However, the control of spacecraft has drawn attention to the fact that minimum fuel consumption may often be more important than minimum settling time. In some cases, the problem of control with minimum fuel consumption leads to a bang-bang control just as in the minimum time control problem. Whether this occurs or not, depends on the type of mechanical or electrical power supply<sup>1</sup>. The following will be restricted to linear systems and to performance criteria which lead to bang-bang control.

Let the system be given by

$$\dot{\vec{x}} = A\vec{x} + B\vec{u} \quad \text{or} \quad \dot{x}^i = f^i(x^j, u) \quad (1)$$

$A$  is a constant matrix and  $B$  is a constant vector. The control function  $u$  is bounded

$$|u| \leq 1 \quad (2)$$

It is desired to go from

$$\vec{x}(0) = \vec{x}_0 \quad \text{to} \quad \vec{x}(T) = \vec{x}_f \quad (3)$$

with the performance criterion

$$\int_0^T g(x^i) dt \rightarrow \text{minimum} \quad (4)$$

Pontryagin's maximum principle states that the Hamiltonian

$$H = \sum p_i \dot{x}^i - g(x^i) \quad (5)$$

must assume at any time an extreme value in order to satisfy the performance criterion. This extremum will be reached, if

$$u = \text{sgn} \sum (p_i b_i) = -\text{sgn} F(t) \quad (6)$$

The functions  $p_i$  are the solutions of a system which is called 'adjoint' to the system considered,

$$\dot{p}_i = -\frac{\partial H}{\partial x^i} = -\sum \frac{\partial f^j}{\partial x^i} p_j \quad (7)$$

The initial conditions of the functions  $p_i$  must be chosen such that starting at  $\bar{x}_0$  the point  $\bar{x}_f$  is reached in time  $T$ . In the minimum time case,  $T$  is not given but is made a minimum, in this case,  $g(x^i) = 1$ . Without specialization  $\bar{x}_f$  can be set  $\bar{x}_f = 0$ .

The above procedure means that to each point  $\bar{x}_0$  belongs a set of switching points given by  $F(t) = 0$ ; see eqn (6). The computation of the coordinates of these switching points for given initial conditions of the phase trajectory is sometimes tedious. If it is possible to find the geometric locus of all possible switching points in a phase space  $[(n-1) \text{ dimensional surface for an } n\text{th order problem}]$ , then the task of finding the initial conditions for the adjoint functions in a particular case is superfluous. For second-order systems without zeros this task has been achieved for the minimum time case<sup>2</sup>, several minimum error criteria<sup>3</sup> and minimum fuel consumption<sup>1</sup>. In the second-order case the locus of the switching points is a curve which separates the phase plane in two halves. The initial  $u$  value is apparent and the only trouble is the realization of this switching curve as a function of the phase variables. The remarkable fact is that the switching curves for minimum time and several error criteria are neighbour curves. This fact first observed for the  $(1/s^2)$  plant has been used by the authors to find the mean square error switching curves for the  $(1/s^2 + 1)$  and  $(1/s^2 + 2\zeta s + 1)$  plants by merely perturbing the minimum time switching curve, and watching the change of the magnitude of the integral determining the performance.

The fact that there exists a mathematical procedure principally to compute the optimum control law, does not mean that the result of such a computation necessarily enables the designer to realize this control. It will first be necessary to weigh the advantage of an optimum control with a possibly difficult control law against a control with a somewhat simpler switching function. Only if the advantage is great will it be decided to realize the optimum law.

The simplest switching function, that is, the linear function, is excluded from this consideration. This switching function will, in general, lead to chatter at least near the origin of the phase space if not also in other regions of the phase space. Optimum control can also lead to chatter. The case of the performance criterion

$$\int_0^T e^2 dt \quad (8)$$

treated by Fuller<sup>3</sup>, shows this clearly, but in general, chatter will be avoided if optimum control is used. In systems with linear switching functions the chatter is due to imperfections, while in systems with optimum control this chatter occurs in an ideal or perfect system and the imperfections of the control components would only modify this chatter.

### Quasi-optimal Switching Curves for Second-order Systems

Linear switching as an approximation of an optimum control function is excluded from this paper<sup>4, 5</sup>. The question of a better approximation, then, has to be dealt with. For performance criteria which do not lead to chatter near the origin, the control law requires a switching curve near the origin which is formed by portions of all possible zero-trajectories. Such a requirement can be and has been easily satisfied, e.g. in the minimum settling time control of second-order systems with pure, imaginary poles. The first example (Figure 1) shows an often expressed idea of approximation for a system with the transfer function  $(1/s^2 + 1)$ <sup>6</sup>. This idea can easily be extended to systems with  $(1/s^2 + 2\zeta s + 1)$ , see Figure 2. In this case the linear part of the switching curve is made parallel to the 'envelope' of the cusps. In both cases one can argue whether much is really lost by taking only part of the first cusp and the dashed lines.

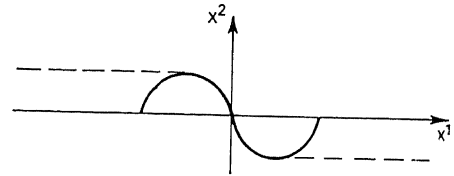


Figure 1. Quasi-optimum switching curve for a  $(1/s^2 + 1)$  system

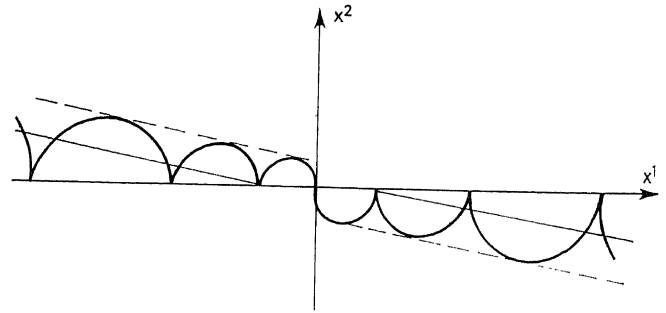


Figure 2. Quasi-optimum switching curve for a  $(1/s^2 + 2\zeta s + 1)$  system

Second-order systems with real poles require switching curves which can be described by a simple power law. Also, the number of switchings required can easily be determined; it is one or none (for all those initial points which incidentally lie on a zero trajectory). The design problem is simple compared with systems with complex or imaginary poles.

### Control of Third-order Systems

For third-order systems the control with minimum settling time is certainly the simplest one. Let the general equation of a system ( $e$  = error = input-output) first be considered

$$e''' + a_2 e'' + a_1 e' + a_0 e = b_2 u'' + b_1 u' + b_0 u \quad (9)$$

This system has poles and zeros. Switching occurs on a surface in the three-dimensional phase space. If  $b_1 = b_2 \equiv 0$ , one has the third-order system without zeros, which will be treated here. The case of systems with zeros can be reduced to the case without zeros by introducing an additional transformation. This case will be treated in a report which will appear in the near future.

In the case of three real poles, the division of the phase space in two halves does not pose any analytical difficulties. The realization of this switching surface may pose an analogue computer problem, but certainly not a digital computer problem. Therefore, the immediate concern is with the problem of one real and two complex poles. This case has been treated recently in two papers<sup>7, 8</sup>. In these papers iteration procedures were described for finding the exact switching curves.

There is no doubt that the exact switching surface poses a difficult design problem, particularly if initial disturbances of any larger size would be admitted. (The magnitude of disturbances should be measured by the bound on the control variable  $u$ . For instance if  $e_0$ ,  $e_0'$  and  $e_0''$  are the initial values of the phase variables, the norm  $\|\vec{e}_0\| \gg |u|_{\max}$  would mean a large disturbance.) Fortunately the experiences with the second-order system can be generalized.

The system described by the following equation is now considered

$$(s + \gamma)(s^2 + 2\zeta\omega s + \omega^2)e = u = -\text{sgn } F, \text{ with } |\zeta| < 1 \quad (10)$$

This third-order differential equation can be replaced by a system of three coupled first-order equations in  $e$ ,  $e'$ , and  $e''$ . This system can then be conveniently transformed to a partially uncoupled system by the transformation

$$\vec{x} = \begin{bmatrix} \gamma & 1 + \frac{\zeta\gamma}{\omega} & \frac{\zeta}{\omega} \\ 0 & \frac{\gamma\gamma}{\omega} & \frac{\gamma}{\omega} \\ \omega^2 & 2\zeta\omega & 1 \end{bmatrix} \vec{e} \quad (11)$$

In the new coordinates the system (10) is described by

$$\dot{\vec{x}} = \begin{bmatrix} -\zeta\omega & \gamma\omega & 0 \\ -\gamma\omega & -\zeta\omega & 0 \\ 0 & 0 & -\gamma \end{bmatrix} \vec{x} + \begin{bmatrix} \zeta/\omega \\ \gamma/\omega \\ 1 \end{bmatrix} u \quad (12)$$

The exact optimum control function  $u$  is given by

$$u = \text{sgn} \left( \frac{\zeta}{\omega} p_1 + \frac{\gamma}{\omega} p_2 + p_3 \right) \quad (13)$$

The functions  $p_i$  are the solutions of the adjoint system

$$\begin{cases} \dot{p}_1 = \zeta\omega p_1 + \gamma\omega p_2 \\ \dot{p}_2 = \zeta\omega p_2 - \gamma\omega p_1 \\ \dot{p}_3 = p_3\gamma \end{cases} \quad (14)$$

Integration of this system yields

$$\begin{cases} p_1 = m_1 e^{\omega\zeta t} \cos(\gamma\omega t + \beta_1) \\ p_2 = -m_1 e^{\omega\zeta t} \sin(\gamma\omega t + \beta_1) \\ p_3 = m_2 e^{\gamma t} \end{cases} \quad (15)$$

with  $m_1$ ,  $m_2$  and  $\beta_1$  as constants of integration. Upon introducing these expressions into eqn (13) one obtains

$$\begin{aligned} u &= \text{sgn} \left[ \frac{\zeta}{\omega} m_1 e^{\omega\zeta t} \cos(\gamma\omega t + \beta_1) \right. \\ &\quad \left. - \frac{\gamma}{\omega} m_1 e^{\omega\zeta t} \sin(\gamma\omega t + \beta_1) + m_2 e^{\gamma t} \right] \\ &= \text{sgn} [m_1^* e^{\omega\zeta t} \cos(\gamma\omega t + \beta_1^*) + m_2 e^{\gamma t}] \end{aligned} \quad (16)$$

The constants  $m_1^*$ ,  $\beta_1^*$  and  $m_2$  depend on the given initial disturbance as mentioned earlier. Their determination causes trouble which can be avoided if a switching surface can be found.

The analytic expression for the exact switching surface is not known. However, we can visualize the form of this surface<sup>7</sup>. This visualization will help us to replace the exact switching surface by a simpler surface. Near the origin of the phase space the approximation of the exact switching surface has to be made more carefully than at larger distances from the origin.

The approximation for larger distances from the origin is developed first and will be called the 'outer' switching surface, then the approximation in the neighbourhood of the origin, called the 'inner' switching surface will be developed.

#### The 'Outer' Switching Surface for Plants $1/s(s^2 + 1)$

In the following a quasi-optimum switching surface for the plant with the transfer function  $1/s(s^2 + 1)$  is first developed and later it is shown how the results for this plant can be generalized.

For  $\gamma = \zeta = 0$  eqn (16) simplifies to

$$u = \text{sgn} [m_1^* \cos(\omega t + \beta_1^*) + m_2] \quad (17)$$

If  $T$  is assumed to be the time for zeroing an initial disturbance, one can consider the study of the motion in reverse time  $\tau$ .

$$\left. \begin{aligned} \tau &= T - t \\ u &= \text{sgn} [m_1^* \cos(\omega\tau - \omega T - \beta_1^*) + m_2] \\ &= \text{sgn} [m_1^* \cos(\omega\tau - \beta_1^{**}) + m_2] \end{aligned} \right\} \quad (18)$$

One can construct a trajectory in reverse time by assuming  $\beta_1^{**}$  and  $(m_1^*/m_2)$ . The switching times are then determined and after  $T$  sec a point in the phase space will be reached which corresponds to the initial disturbance. Since  $\cos(\omega\tau - \omega T - \beta_1^*) = \cos[\omega\tau - (\omega T + \beta_1^*)]$  is a periodic function, two trajectories for which  $(\omega T_1 + \beta_{11}^*) - (\omega T_2 + \beta_{12}^*) = 2\pi n$  will partially coincide and have coinciding switching points in the identical portions. It has been shown<sup>7, 8</sup> how to find the

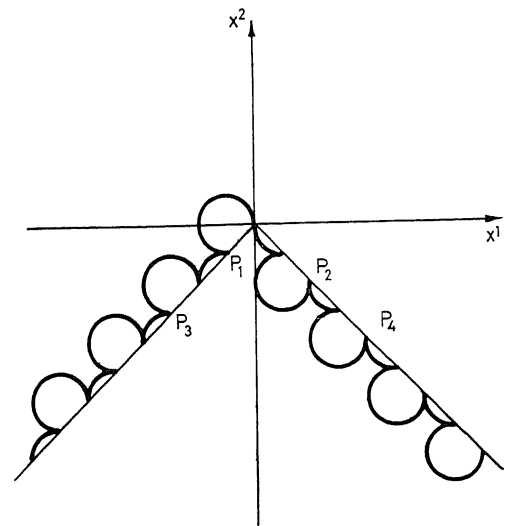


Figure 3. Projection of the optimum switching curve in the  $x_1 x_2$  plane for a  $(1/s(s^2 + 1))$  system

location of these switching points in the phase space (see Figure 3). It is obvious that it would be difficult to build-up the surface on which all possible switching points for all possible  $\beta_1$ ,  $m_1^*/m_2$ ,  $T$  are lying, but there is no need to be very exact as soon as one is one cusp away from the origin of the phase space. Therefore it was tried to approximate the locus of the switching points in a rather primitive way. The straight lines on which the

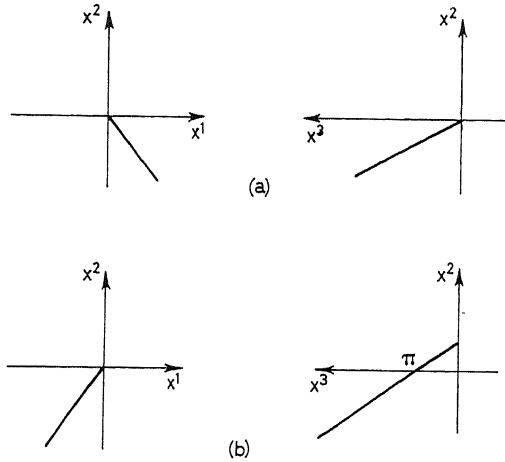


Figure 4. Projections of lines joining the tips of the cusps in the  $x^1 x^2$  and  $x^2 x^3$  planes respectively

points  $P_2, P_4 \dots$  [Figure 4 (a)] and  $P_1, P_3 \dots$  [Figure 4 (b)] are lying, are considered as representatives of the more complicated curve which is the carrier of switching points. If one considers the ruled surface built of these lines, one obtains for  $\|x\| > 1$ .

$$u = -\operatorname{sgn} F$$

$$= -\operatorname{sgn} \left\{ 2|x^1| x^3 + (\operatorname{sgn} x^2) [(x^1)^2 + (x^2)^2] \right. \\ \left. \times \arccos \left[ \frac{(x^1)^2 - (x^2)^2}{(x^1)^2 + (x^2)^2} \right] \right\} \quad \text{for } (x^1 x^3) > 0 \quad (19a)$$

and

$$u = -\operatorname{sgn} \left\{ 2|x^1| [x^3 - (\operatorname{sgn} x^3) \cdot \pi] + (\operatorname{sgn} x^2) [(x^1)^2 + (x^2)^2] \right. \\ \left. \times \arccos \left[ \frac{(x^1)^2 - (x^2)^2}{(x^1)^2 + (x^2)^2} \right] \right\} \quad \text{for } (x^1 x^3) < 0 \quad (19b)$$

Figure 5 shows a sketch of the surface which includes the  $x^3$  axis. On  $x^3$ , however, there do not lie real switching points though  $x^1 = x^2 \equiv 0$  yields  $F = 0$ . This is because there is no change of sign. The function  $F$  behaves as indicated in Figure 6.

#### The 'Inner' Switching Surface

Mention must now be made of the switching surface close to the origin. The visualization is easy. In reverse time the phase point leaves the origin on a zero trajectory, each point of which can be considered as a switching point. That means, from each point of the two zero trajectories a new trajectory is emerging. These new trajectories form a surface. Each point of this surface can be considered again as a potential switching point, which means as a starting point of a new trajectory portion in reverse

time. As indicated earlier only the surface formed by the trajectories emanating from the two zero trajectories will be considered. The portion of this surface, for which the distance of the surface points from the origin is smaller than a fixed value, will be used. For larger distances the surface given by eqn (19) will serve.

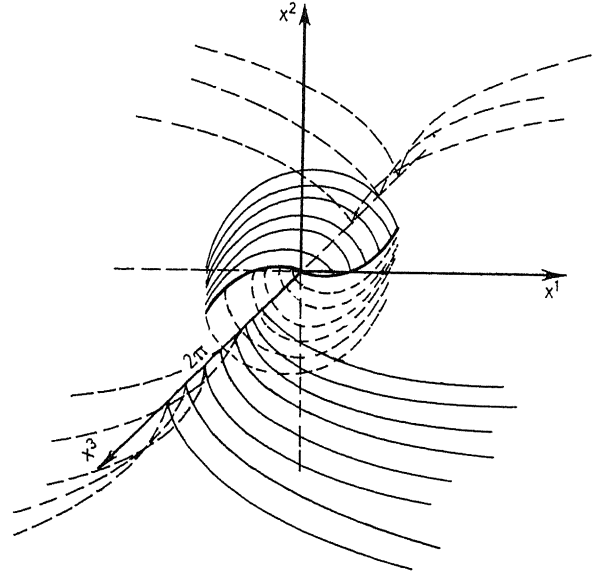


Figure 5. Sketch of the quasi-optimum switching surface for a  $(1/s(s^2 + 1))$  system

The limit for the use of the surface near the origin was first assumed to be given by  $[(x^1)^2 + (x^2)^2 + (x^3)^2] < (2)^2$ . However, it soon turned out that the transition from the 'outer' surface to the 'inner' surface can cause trouble. Since the outer surface is

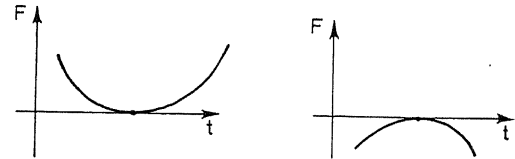


Figure 6. Sketch of the behaviour of  $F$  vs time as the  $x^3$  axis

not the exact switching surface, the phase point may pass, let us say, a switching from  $(+1)$  to  $(-1)$ , just before transition. However, this switching may bring the phase point to the  $(+1)$  side of the inner surface. Therefore a contradiction exists and a chatter occurs which causes the phase point to be trapped in the transition region. This trouble can usually be avoided by taking  $[(x^1)^2 + (x^2)^2 + (x^3)^2] < 1$ . Because of the flatness of the outer surface near the origin, it is even better to use  $[(x^1)^2 + (x^2)^2] < 1$  as the transition condition.

The analytic determination of the switching surface close to the origin (the 'inner' surface) is somewhat troublesome, even if one considers its realization by digital or analogue computer equipment. Therefore some simplifications are desired. Two possibilities are available.

*First possibility:* One replaces the plant transfer function

$$\frac{1}{(s+\gamma)(s^2+2\zeta s+1)} \quad \text{by} \quad \frac{1}{s^3} \quad (20)$$

For the latter transfer function, the switching surface is given by

$$F = e_1 + \frac{1}{3}e_3^3 + we_3e_2 + w\left(\frac{1}{2}e_3^2 + we_2\right)^{3/2} = 0 \quad (21)$$

with  $e_1 = e$ ,  $e_2 = \dot{e}$ , and  $e_3 = \ddot{e}$ . The factor  $w$  is determined as

$$\text{and } w = \begin{cases} +1 & \text{for } \left[ e_2 + \frac{1}{2}e_3|e_3| \right] > 0 \\ -1 & \text{for } \left[ e_2 + \frac{1}{2}e_3|e_3| \right] < 0 \end{cases} \quad (21a)$$

From eqn (21) the control function is obtained for the neighbourhood of the origin

$$u = -\operatorname{sgn} F = -\operatorname{sgn} \left[ e_1 + \frac{1}{3}e_3^3 + we_3e_2 + w\left(\frac{1}{2}e_3^2 + we_2\right)^{1/2} \right] \quad (22)$$

with the same rule for the signs.

The deviations caused by using the model  $(1/s^3)$  instead of the correct plant can be visualized in the following figures.

In Figure 7 the zero trajectories are given in the  $e_1 e_2$  plane for a full third-order system and three possible approximations to it. The output of the relay is either  $(+1)$  or  $(-1)$ . This determines the scale of the figures. Naturally one can have more and less agreement depending on the values of  $\gamma$  and  $\zeta$  which here are unit and zero respectively.

In Figure 8 projections of these trajectories into the  $e_2 e_3$  plane are shown. It can be clearly seen that only in a rather limited region are the zero trajectories of the approximations close to the zero trajectory of the original system.

In Figure 9 projections of zero trajectories for several other third-order systems are shown. If the initial values are not too large and only one switching occurs between the start and the reaching of the origin of the phase space, these curves will give an idea of how good the approximation of the optimum control

will be, if the complete third-order system is replaced by simpler ones.

The second possibility for control near the origin of the phase space is modification of the surface given by eqn (19). One applies factor  $N$  to the control function which diminishes the control effort; that means

$$u = -\operatorname{sgn} F \quad (23a)$$

will be replaced by

$$u = -N \operatorname{sgn} F \quad (23b)$$

with  $N = f(\|x\|)$  as indicated in Figure 10. A step width  $\Delta$  has to be chosen and  $N_i = (i+1)\Delta$  for  $\|x\| = i\Delta + \varepsilon$  with  $0 < \varepsilon < \Delta$ . In this procedure, one may say, the neighbourhood of the origin is stretched. Naturally one cannot expect to reach the absolute 'zero', the final state will be a chatter around the origin. Also, one has to count on losing some time by diminishing the control effort.

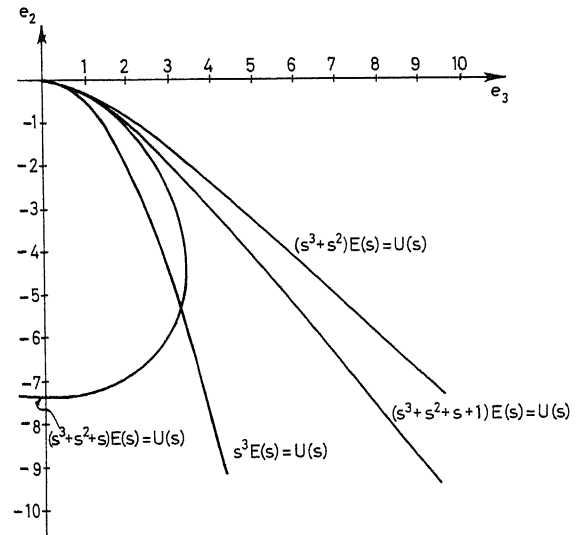


Figure 8. Projection of the zero trajectories in the  $e_2 e_3$  plane for a full third-order system and three approximations to it

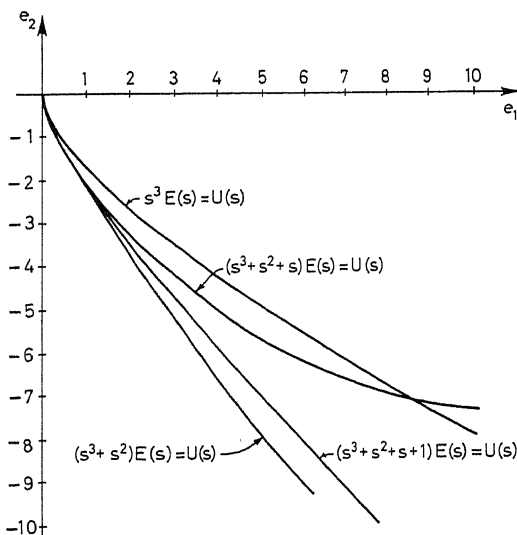


Figure 7. Projection of the zero trajectories in the  $e_1 e_2$  plane for a full third-order system and three approximations to it

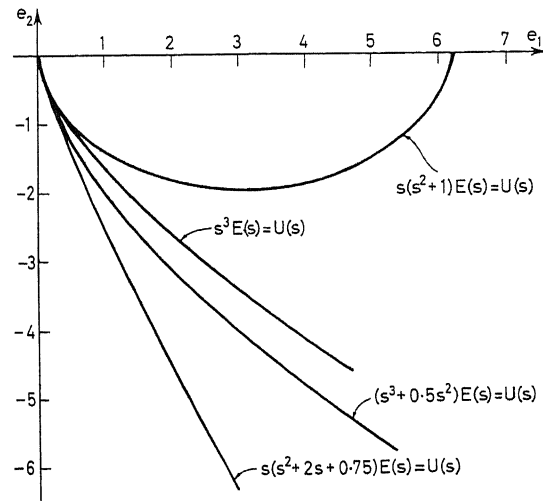


Figure 9. Projection of the zero trajectories for several third-order systems

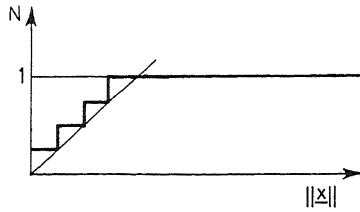


Figure 10. Sketch of multi-level control behaviour near the origin of the phase space

### Results

The eqns (19a, b) and (22) for the switching surface may look rather complicated, but they can easily be implemented with the help of a modern computer. It is expected that new miniature digital computing elements can be used in spacecraft.

A number of examples have been investigated. These examples were first constructed in reverse time by employing the exact switching points, then it was assumed that the initial disturbance was given and that the 'approximate' switching surfaces given by eqns (19) and (22) were used to zero this disturbance. One example is given here.

$$e_1^0 = 30.67; \quad e_2^0 = 2.93; \quad e_3^0 = -7.07$$

True optimum and quasi-optimum switching time were practically the same. In Figures 11(a) and (b) projections of the phase trajectories are shown. Two other examples are given here

$$\left. \begin{array}{l} e_1^0 = 13 \\ e_2^0 = 13 \\ e_3^0 = 2 \end{array} \right\} \text{ i.e. } \left\{ \begin{array}{l} x_0^1 = 13 \\ x_0^2 = 2 \\ x_0^3 = 15 \end{array} \right.$$

The optimum time for this example is  $T_{\text{opt}} = 31.3$ , and the quasi-optimum time is  $T_{q.o.} = 32.4$ , at which time  $|x_j^i| \leq 0.1$ .

For the next example, with

$$\left. \begin{array}{l} e_1^0 = 6.63 \\ e_2^0 = e_3^0 = 0 \end{array} \right\} \text{ i.e. } \left\{ \begin{array}{l} x_0^1 = 0 \\ x_0^2 = 0 \\ x_0^3 = 6.63 \end{array} \right.$$

$T_{\text{opt}} = 8.3$  and  $T_{q.o.} = 8.9$  was obtained with  $|x_j^i| \leq 0.1$  for the latter one.

### Outer Switching Surface for Plants with a More General Transfer Function of Third Order

As indicated earlier, there is still the task of developing a quasi-optimum outer switching surface for plants with  $\gamma \neq 0$  and  $\zeta \neq 0$ . Consider first the case when  $\gamma = 0$  and  $\zeta \neq 0$ . When the analytical expressions for the canonical variables as functions of time are compared, it is recognized that for  $\zeta \neq 0$  in the  $n$ th interval\*

$$x_n^2 = C_n e^{-\zeta x_n^3} \cos(\nu x_n^3 + \delta_n) \quad (24a)$$

with  $\nu = (1 - \zeta^2)^{1/2}$  compared with

$$x_n^2 = C_n^* \cos(\nu x_n^3 + \delta_n^*) \quad (24b)$$

for  $\zeta = 0$ .  $x_n^1$  is similarly changed. Therefore it is indicated that eqns (19) should change correspondingly.

\* See eqns 12 (a) and 12 (b) with  $\omega = 1$ . Replace  $t$  by  $x^3$  with the help of the third equation  $\dot{x}^3 = -\gamma x^3 + u$  or  $\dot{x}^3 = u$  for  $\gamma = 0$ .

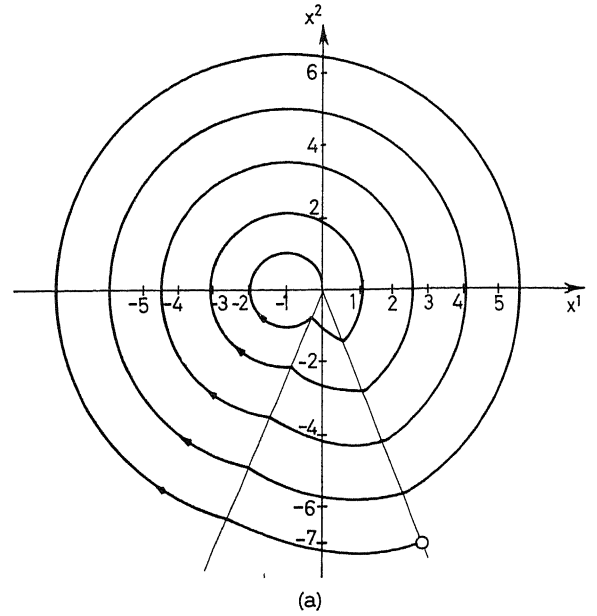


Figure 11 (a). Projection of the optimum trajectory in the  $x^1 x^2$  plane for a system with  $\gamma = 0$ ,  $\zeta = 0$ . Initial disturbance  $x^1 = e_2 = 2.93$ ,  $x^2 = e_3 = -7.07$ ,  $x^3 = 7.5\pi$ , i.e.  $e_1 = x^2 + x^3 = 30.67$

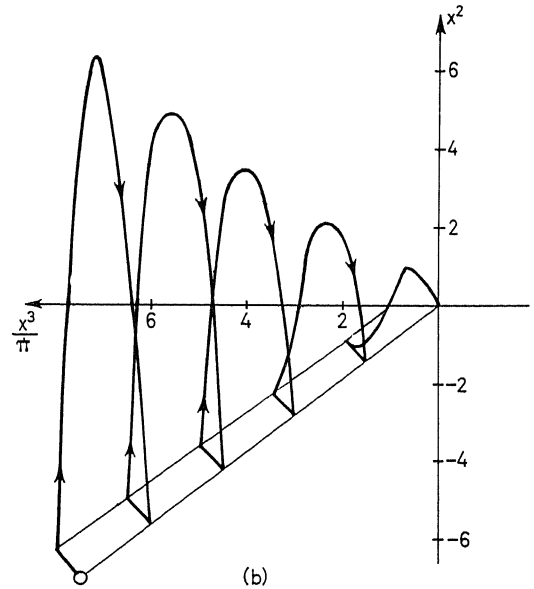


Figure 11 (b). Projection of the same trajectory in the  $x^2 x^3$  plane

$$u = -\text{sgn} \left\{ 2|x^1| x^3 e^{+\zeta|x^3|} + (\text{sgn } x^2) [(x^1)^2 + (x^2)^2] \right. \\ \left. \times \arccos \left[ \frac{(x^1)^2 - (x^2)^2}{(x^1)^2 + (x^2)^2} \right] \right\} \quad \text{for } (x^1 x^3) > 0 \quad (25a)$$

and

$$u = -\text{sgn} \left\{ 2x^1 [x^3 - (\text{sgn } x^3)\pi] e^{+\zeta|x^3|} \right. \\ \left. + (\text{sgn } x^2) [(x^1)^2 + (x^2)^2] \arccos \left[ \frac{(x^1)^2 - (x^2)^2}{(x^1)^2 + (x^2)^2} \right] \right\} \\ \text{for } (x^1 x^3) < 0 \quad (25b)$$



This is a somewhat primitive generalization, because a look at the figures in (7) shows that one could apply a much more strict analysis. For instance, one could build-up the switching surface out of the straight lines which are the arithmetic means between the lines  $S_2', S_4', \dots$  and  $S_1, S_3, \dots$  (see Figure 4)<sup>7</sup>. In the limiting case ( $\zeta \rightarrow 0$ ) this new line would converge into the old one.

The situation becomes more involved, when  $\gamma \neq 0$ . In this case

$$x_n^3 = \frac{u_n}{\gamma} + \left( x_{n0}^3 - \frac{u_n}{\gamma} \right) e^{-\gamma t_n} \quad (26a)$$

compared with

$$x_n^3 = u_n t_n + C_n^* \quad \text{for } \gamma = 0 \quad (26b)$$

Consider for brevity, the case with  $\zeta = 0$ . One sees immediately that the line connecting points  $0, P_2, P_4, \dots, P_{2n}$  is no longer a straight line, in spite of the fact that the projection into the  $x^1 x^2$  plane may still be assumed to be straight. This assumption is in the frame of the approximation described by Flüge-Lotz<sup>7, 8</sup>. It is possible to establish a difference equation for

$$\frac{x^3(P_{2n}) - x^3(P_{2n-2})}{x^2(P_{2n}) - x^2(P_{2n-2})} \quad (27)$$

which leads to a differential equation for the relation  $x^3(x^2)$ . The quasi-optimum switching surfaces  $F(x^1, x^2, x^3) = 0$  then can be found. It can be realized easily with digital equipment. In the near future it is intended to increase the number of examples and to compare more quasi-optimum results obtained with the truly optimal results. Also, the results are being extended to include the general case with finite  $\gamma$  and finite  $\zeta$ .

#### A Simple Fourth-order Problem

The importance of two-axes satellite control directed attention to a fourth-order problem, described by

$$(s^2 + 2\zeta_1\omega_1 s + \omega_1^2)(s^2 + 2\zeta_2\omega_2 s + \omega_2^2)e(s) = u(s) \quad (28)$$

In orthogonal variables the system is given by four differential equations of first order.

$$\dot{\bar{x}} = \begin{bmatrix} -\zeta_1\omega_1 & v_1\omega_1 & 0 & 0 \\ -v_1\omega_1 & -\zeta_1\omega_1 & 0 & 0 \\ 0 & 0 & -\zeta_2\omega_2 & v_2\omega_2 \\ 0 & 0 & -v_2\omega_2 & -\zeta_2\omega_2 \end{bmatrix} \bar{x} + \begin{bmatrix} \zeta_1/\omega_1 \\ v_1/\omega_1 \\ \zeta_2/\omega_2 \\ v_2/\omega_2 \end{bmatrix} u \quad (29)$$

The control torque depends only on one control function; this clearly is a simplification for a first attack of the problem. The case of two steering functions is also being studied. The homogeneous system is decoupled in two systems of second order. However, the optimal control function depends on both frequencies (example  $\zeta = 0$ )

$$u = \text{sgn} [C_1 \cos(\omega_1 \tau + \delta_1^*) + C_2 \cos(\omega_2 \tau + \delta_2^*)] \quad (30)$$

where  $\tau = T - t$  denotes again reverse time. Therefore the two second-order systems are coupled through the control function  $u$ .

A representation of the phase trajectory can only be made by tracing projections into the  $x^1 x^2$  and the  $x^3 x^4$  planes. An

example with  $\omega_1/\omega_2 = 1/3$  serves to acquaint one with the problem. In this case the coordinates in the  $i$ th interval are given by

$$x_i^1 = \frac{u}{\omega_1^2} + R_{10} \cos(\omega_1 t_i + \delta_{1i})$$

$$x_i^3 = \frac{u}{\omega_2^2} + R_{20} \cos(\omega_2 t_i + \delta_{2i})$$

It is obvious that the control-force influence in the  $x^1 x^2$  plane is nine times larger than in the  $x^3 x^4$  plane. Figures 12(a) and (b) show an example, which was designed in reverse time. One recognizes also that in this case an approximation of the time optimal control law is rather easily possible. This example is particularly simple in that the ratio  $\omega_2/\omega_1$  is an integer. Additional, more general examples are being studied.

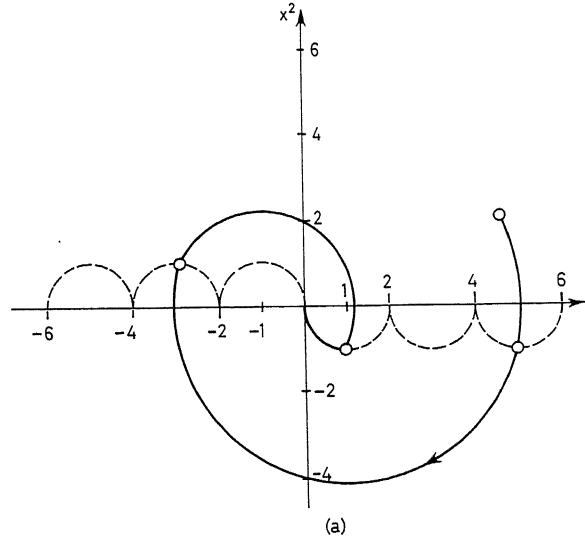


Figure 12 (a). Fourth-order system, projection of the optimum trajectory into  $x^1 x^2$  plane

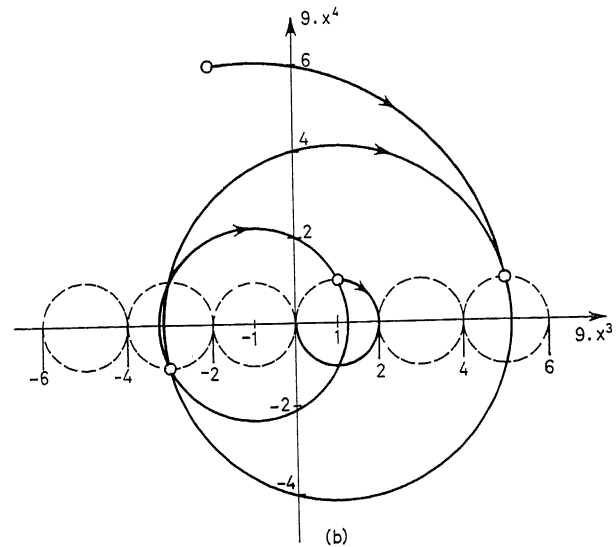


Figure 12 (b). Fourth-order system, projection of the optimum trajectory into the  $x^3 x^4$  plane

## References

- <sup>1</sup> FLÜGGE-LOTZ, I. and MARBACH, H. The optimal control of some attitude control systems for different performance criteria. Paper presented at the JACC 1962 New York; in the press (*Trans. Amer. Soc. mech. Engrs*)
- <sup>2</sup> BUSHAW, D. *Optimal Discontinuous Forcing Terms, Contributions to the Theory of Non-linear Oscillations*, vol. IV, pp. 29-52. S. Lefschetz (Ed.) 1958. Princeton; Princeton University Press
- <sup>3</sup> FULLER, A. T. Relay control systems optimized for various performance criteria. *Automatic and Remote Control*. 1961. London; Butterworths
- <sup>4</sup> FLÜGGE-LOTZ, I. and ISHIKAWA, T. Investigation of third-order contactor control systems with two complex poles without zeros. NASA TN D-428 (1960)
- <sup>5</sup> FLÜGGE-LOTZ, I. and ISHIKAWA, T. Investigation of third-order contractor control systems with zeros in their transfer functions. NASA TN D-719 (1961)
- <sup>6</sup> KNUDSEN, H. K. Maximum effort control for oscillatory elements. *Inst. Radio Engrs N. Y. Wescon Con. Rec.* Pt IV (1959), 116
- <sup>7</sup> FLÜGGE-LOTZ, I. and MIH YIN. The optimum response of second-order, velocity-controlled systems with contactor control. *Trans. Amer. Soc. mech. Engrs, Ser. D, J. Basic Engng*, March (1961), 59
- <sup>8</sup> FLÜGGE-LOTZ, I. and TITUS, H. A. The optimum response of full third-order systems with contactor control. *Trans. Amer. Soc. mech. Engrs, Ser. D, J. Basic Engng*, Dec. (1962) 554

## DISCUSSION

M. HAMZA, *ETH, Zurich, Switzerland*

The work of Professor Flügge-Lotz on optimum control is well known and is estimated highly. The present paper is very useful in that it presents means of approximating the switching surface obtained for optimum control, thus overcoming a major practical disadvantage, that is, simulation of a complex surface.

The results in this paper are given for a system having a fixed structure. If the structure is variable this problem becomes much more complex since a variable switching surface is required. What approach would the authors suggest for realizing a variable switching surface? Has the sensitivity of the switching surface to parameter changes been investigated? The authors did not mention anything about the error obtained using their method, would Professor Flügge-Lotz please comment on this point?

I. FLÜGGE-LOTZ, *in reply*

May I answer, firstly, the third question of Dr. Hamza. In the past year an extensive investigation of the quality of the proposed approximations to the optimal switching surface was made (with assistance of Mr. Y. Kashiwagi after Dr. Titus had left Stanford University). This was done in the usual way: a number of exact time-optimal solutions were computed in backward time. They start at the origin of the error state space and lead, after a chosen time  $T$ , to some state with finite error and error derivatives. Then the latter ones were used

as the initial conditions for the sub-optimal control in forward time. In general, the sub-optimal control will not zero any error derivatives in finite time because the inner switching surface is also an approximation to the exact one. Therefore times for reaching  $|\vec{t}| = 1$  ( $1 =$  a very small number compared to  $|u|_{\max}$ ) with optimal controls were compared. The following results were obtained:

For  $\gamma = 0$  and  $\zeta = 0$ , where the approximating surfaces come closest to the exact switching surface, the sub-optimal solution is very good. For  $\gamma \neq 0$  and  $\zeta \neq 0$  it is suggested that the approximate surfaces be used only for  $\gamma < 0.5$  and  $\zeta < 0.3$ . Then the deviation of the optimal and sub-optimal time for reaching 1 can be kept below 20 per cent. In many cases it will be lower. The choice of the surface on which transition of the 'outer' to the 'inner' switching surface should take place influences the results: other surfaces than those mentioned in the text have been found to be preferable.

The first two questions of Dr. Hamza are related to the above-mentioned facts. If the structure of the system changes slowly (that means  $\gamma$  and  $\zeta$  change slowly) one can vary the approximate switching functions correspondingly and hope to obtain a good sub-optimal solution. A digital programme for realizing this could be established. However, how good this solution is (that means how close it comes to the exact optimal solution), would require exact optimal solutions for such cases to be studied first. I thank Dr. Hamza for his questions which gave me an opportunity to report about the later development of our work.

# Programme Control and the Theory of Optimal Systems

YE. A. BARBASHIN

## Summary

In this paper consideration is given to the system of differential equations

$$\frac{dx}{dt} = f(x, \eta, t) + u(c, y, t)$$

where  $x(t)$  is an  $n$ -dimensional vector,  $y(t)$  an  $m$ -dimensional vector,  $\eta(t)$  a certain (in general random) vector function, and  $c$  a constant vector. It is assumed that a certain trajectory  $x = \psi(t)$  in phase space is given for  $0 \leq t \leq T$  ( $0 < T \leq \infty$ ). Assuming that certain information is received on the variation of  $\eta(t)$ , it is required to choose a vector  $c$  (problem A), or a vector function  $y(t)$  (problem B), or a vector  $c$  and a function  $y(t)$  (problem C), such that some solution of the system precisely or approximately realizes a motion along the trajectory  $x = \psi(t)$ .

The paper shows the methods of programme control used.

## Sommaire

Ce rapport prend en considération le système d'équations différentielles suivant:

$$\frac{dx}{dt} = f(x, \eta, t) + u(c, y, t)$$

dans lequel  $x(t)$  est un vecteur à  $n$ -dimensions,  $y(t)$  un vecteur à  $m$ -dimensions,  $\eta(t)$  une certaine fonction vectorielle en général de caractère aléatoire et  $c$  un vecteur constant. On admet qu'une certaine trajectoire  $x = \psi(t)$  est donnée dans l'espace de phase pour l'intervalle de temps  $0 \leq t \leq T$  ( $0 < T \leq \infty$ ). Compte tenu que certaines informations sont reçues concernant la variation de  $\eta(t)$ , il s'agit de déterminer le vecteur  $c$  (problème A), ou la fonction vectorielle  $y(t)$  (problème B), ou encore  $c$  et  $y(t)$  simultanément (problème C), en sorte que la solution du système s'identifie ou du moins se rapproche de la trajectoire  $x = \psi(t)$ . Le rapport montre les méthodes de commande à programme à utiliser pour résoudre ce problème.

## Zusammenfassung

Der Aufsatz behandelt das System von Differentialgleichungen

$$\frac{dx}{dt} = f(x, \eta, t) + u(c, y, t)$$

wobei  $x(t)$  ein  $n$ -dimensionaler Vektor,  $y(t)$  ein  $m$ -dimensionaler Vektor,  $\eta(t)$  eine bestimmte (im allgemeinen regellose) Vektorfunktion und  $c$  ein konstanter Vektor ist. Dabei wird angenommen, daß eine bestimmte Trajektorie  $x = \psi(t)$  im Phasenraum für  $0 \leq t \leq T$  ( $0 < T \leq \infty$ ) gegeben ist. Angenommen, daß man aus der Variation von  $\eta(t)$  eine bestimmte Information erhält, so ist es erforderlich, einen Vektor  $c$  (Problem A) oder eine Vektorfunktion  $y(t)$  (Problem B) oder einen Vektor  $c$  und eine Vektorfunktion  $y(t)$  (Problem C) so zu wählen, daß eine Lösung des Systems genau oder angenähert die Bewegung entlang der Trajektorie  $\psi(t)$  darstellt. Die Arbeit enthält die Prinzipien der verwendeten Programmsteuerung.

## Introduction

Consideration is given to the system of differential equations

$$\frac{dx}{dt} = f(x, \eta, t) + u(c, y, t) \quad (1)$$

where  $x(t)$  is an  $n$ -dimensional vector,  $y(t)$  an  $m$ -dimensional vector,  $\eta(t)$  a certain (in general random) vector function, and  $c$

a constant vector. It is assumed that a certain trajectory  $x = \psi(t)$  in phase space is given for  $0 \leq t \leq T$  ( $0 < T \leq \infty$ ). Assuming that certain information is received on the variation of  $\eta(t)$ , it is required to choose a vector  $c$  (problem A), or a vector function  $y(t)$  (problem B), or a vector  $c$  and a function  $y(t)$  (problem C), such that some solution of the system precisely or approximately realizes a motion along the trajectory  $x = \psi(t)$ . The problem formulated in this manner is a problem of programme control.

Let  $O$  in Figure 1 be the plant under control, whose object is to achieve a certain given mode of operation  $x = \psi(t)$ . To achieve this, a unit  $Y$  is introduced that develops a control  $u$ . In forming the control, use is made of information on the operating conditions  $\psi(t)$  to be set up and also on external influences  $\eta(t)$ ; this information may be received in distorted form for many reasons, e.g. delays and inertia in the transmission line  $C$ , measurement errors, random errors, etc.

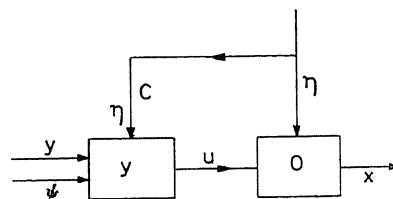


Figure 1

If the problem had an accurate solution, then the required control would be determined by the relation

$$u(c, y(t), t) = \psi'(t) - f(\psi(t), \eta(t), t) \quad (2)$$

However, in a number of cases the system (2) cannot be solved for the control vector  $c$  or the control function  $y(t)$ . This may be due to the choice of an inadequate number of dimensions for the vectors  $c$  and  $y(t)$ , or the presence of incomplete or distorted information on the external influences  $\eta(t)$ . It may also happen that it is possible only to choose the control from some narrowly defined class of functions, such as piecewise-continuous functions, trigonometrical polynomials, functions whose modulus has a constant limit, etc.

Thus the impossibility of solving system (2) accurately may lead to the statement of a number of problems of variational type. Bellman<sup>1</sup> considered such a problem when he used the dynamic programming technique to derive a control function  $y(t)$  so as to make the maximum deviation of the system (1) trajectory from the required one a minimum.

One can state the problem of finding control functions that make a certain integral criterion of control quality a minimum. These problems may be solved by making use of the maximum principle of Pontryagin<sup>2, 3</sup>, of classical variational principles<sup>4</sup>, and of the principle of dynamic programming<sup>5</sup>.

One can state the problem of minimizing the error with which the given trajectory satisfies system (1). If it is a question

of a minimum mean-square error, then this statement of the problem leads to the simplest problems in the theory of mean-square approximations<sup>6</sup>.

Finally, one can renounce all attempts to minimize the deviation, and seek a solution that simply gives a sufficiently accurate approximation. Thus, for example, Roytenberg<sup>7</sup> seeks a control function from the class of piecewise-continuous functions to give coincidence with the system (1) trajectory at a finite number of points on it.

It is observed that a distinction should be made between two essentially different cases in solving the problem of realizing the given process. In the first case the initial points of the actual and the required trajectories coincide; in the second the initial condition of the actual process may have any value. Normally in the second case the control is formed not only according to the magnitude of the disturbance but also according to the deviation of the controlled quantity from that required, i.e. in this case the control system will include feedback.

### Programme Control and Optimal Systems

It should be noted that if the optimum principle is satisfied in one form or another in solving the problem of realizing a motion along a given trajectory, then in a number of cases it becomes possible to introduce feedback into the control system, which permits one to automatically correct the motion along that trajectory. This fact proves conclusively the advantage of systems designed on the basis of one or other optimum criterion. However, a number of difficulties are met in applying optimum criteria to the design of programme control systems and tracking systems. First of all, they often lead to designs requiring heavy computation or to designs that are technically difficult or impossible to execute. In this case, one must give up trying to satisfy the optimum principle, and restrict oneself to an approximate solution of the problem, with the aim of getting the best quality of fit within the bounds of technical possibility.

Usually in applying the optimum principle one needs a knowledge of the process to be realized over its whole duration. But it may happen that only certain statistical characteristics of the programmed process are known, or even that nothing is known about its future. In the latter case the optimal control theory is powerless, and the only reasonable approach is that of minimizing the deviation of the velocity vector for the current point from the tangent vector to a certain curve of pursuit from a given class, perhaps determined by a system of differential equations. Thus in this case the problem of minimizing the deviation of the given process from that required is replaced by the problem of minimizing the difference between two vector fields, one of which determines the actual motion of the point and the other the required motion along a curve of pursuit. It should be noted that an analogous result is obtained if the control is chosen to give a maximum rate of decrease of a certain Liapunov function set up for the perturbed-motion system. The above approach, by introducing feedback, enables the deviation of the actual process to be rapidly reduced from that required, without the future course of the latter being known. It is shown below that an analogous result can be obtained by increasing the stability of the basic control circuit.

If the motion to be realized is known over the whole of its duration, then the following is the most natural method of solving the programme control problem. In the first stage a

control that gives the most rapid means of reaching the given trajectory is sought, and in the second the control that achieves motion along that trajectory<sup>8</sup> is found.

The mathematical theory of optimal control that exists at present is basically a theory of optimal stabilization. This means that this theory permits, in the simplest cases, by the introductions of relay devices into the control system, an increase in the system's closed-circuit stability at zero input signal. In other words, the quality of the system's stability is improved, using optimum criteria, irrespective of the nature and type of input actions that are processed. Clearly such a system will deal with input actions in various manners according to their structure. Below is given an example of a servo-system that reacts well enough to step-function inputs, and consequently also to any slowly varying inputs. However, in order to obtain this good quality in the system, the requirement for optimum closed-circuit stability had to be abandoned.

### Example

The control system having the block diagram shown in Figure 2 is considered. Here  $f$  is the input signal,  $x$  the output signal,  $K$  the amplifier unit,  $A$  the unit forming the gain  $\kappa$ , which in general is variable. The problem is to find the optimum law of variation, given the constraint  $|\kappa| \leq \kappa_0$ , and the condition that the error  $\varepsilon$  decreases in some sense in the fastest way.

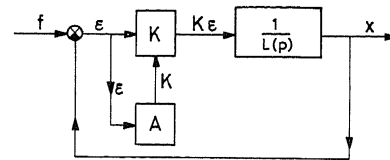


Figure 2

Since in this case the time for complete elimination of the error is infinite from a mathematical point of view, the time for the error to fall within a given region surrounding the origin has to be discussed. Rapid action of this type will be called relative rapid action in distinction to the normal type.

The case where the plant under control is specified by an equation of the second order is considered, i.e. where  $L(p) = p^2 + ap + b$ . In this case the differential equation being examined takes the form

$$\ddot{\varepsilon} + a\dot{\varepsilon} + b\varepsilon = \ddot{f} + a\dot{f} + bf - \kappa\varepsilon \quad (3)$$

First consider the case where the external input  $f$  is absent, and find the law of variation of  $\kappa$  that gives relative rapid action. According to the results of Yemelyanov and Fedotova<sup>9</sup>, the gain should be determined by the formula

$$\kappa = \kappa_0 \operatorname{sgn} \varepsilon (Tp + 1) \varepsilon \quad (4)$$

where  $T^{-1}$  is the negative root of the equation

$$\lambda^2 + a\lambda + b - \kappa_0 = 0 \quad (5)$$

Now consider the case where  $f$  is a step function, i.e. let  $f = 0$  for  $t < 0$  and  $f = f_0$  for  $t \geq 0$ , assuming that the magnitude  $f_0$  of the step cannot be measured. Reserving the freedom to choose  $T$ , take the previous switching law given by eqn (4).

Clearly if  $\kappa = \kappa_0$ , eqn (3) after the step has taken place will have the form

$$\ddot{\varepsilon} + a\dot{\varepsilon} + (b + \kappa_0)(\varepsilon - \varepsilon_2) = 0 \quad (6)$$

where  $\varepsilon_2 = bf_0/(b + \kappa_0)$ .

Correspondingly, for  $\kappa = -\kappa_0$  one gets

$$\ddot{\varepsilon} + a\dot{\varepsilon} + (b - \kappa_0)(\varepsilon - \varepsilon_1) = 0 \quad (7)$$

where  $\varepsilon_1 = bf_0/(b - \kappa_0)$ .

Assuming that  $\kappa_0$  is large enough a qualitative plot in the phase plane can be drawn for each of these equations without difficulty. Equation (6) in the phase plane corresponds to a family of spirals converging to a focus-type special point  $(\varepsilon_2, 0)$ . Equation (7) in the phase plane corresponds to a family of integral curves of hyperbolic type, with a 'saddle'-type special point  $(\varepsilon_1, 0)$  through which pass two integral straight lines whose gradients are the roots of eqn (5).

Assuming now that the switching law is given by eqn (4), the phase diagram shown in Figure 3 is obtained, provided only that  $T\lambda_1 < -1$ , where  $\lambda_1$  is the negative root of eqn (5), is assumed. If the latter inequality is not satisfied, an obviously unsatisfactory result is arrived at, since the switching line (T)

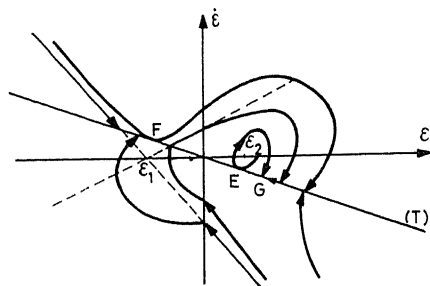


Figure 3

given by the equation  $T\dot{\varepsilon} + \varepsilon = 0$  will be cut by the integral curves over the whole of its length with  $\varepsilon < 0$ , while in our case the straight line T is a sliding line everywhere except over the segment EF, where E and F are the points of contact with the integral curves corresponding to eqns (6) and (7). Thus the switching line resulting from the relevant optimum criteria will have been deliberately abandoned. If the representative point M falls to the left of the line T, then it will slide along this line as far as F, follow a curve of hyperbolic type as far as the line  $\varepsilon = 0$ , then a spiral as far as the right-hand part of line T, where it will again start to slide towards E. On arriving at E it will approach the point  $(\varepsilon_2, 0)$  along a spiral if  $a > 0$ , while if  $a < 0$  it will start to move along a cycle consisting of the segment GE of line T and the segment EHG of the spiral. Thus any point in the plane arrives, eventually, either within a sufficiently small region about the point  $(\varepsilon_2, 0)$  or at a limit cycle corresponding to some self-oscillatory mode. It should be observed that the amplitude of the resulting self-oscillations is of the same order as  $\varepsilon_2 = bf_0/(b + \kappa_0)$ , and consequently can be made as small as required by increasing  $\kappa_0$ .

It should be noted that by increasing T the length of the segment over which it cuts the integral curves is decreased, but the speed of sliding along this line is also lessened since, as can readily be seen, the sliding law is given by the relation  $\varepsilon = \varepsilon_0 \exp(-t/T)$ . Thus proceeding from various quality criteria and combining speculation with experiment a reasonable value for the time constant T can be selected.

Together with R. M. Yeydinov and V. A. Tabueva the author has been carrying out analogous investigations for a third-order system. Here the main difficulty lies in the problem of synthesizing a corresponding optimal system.

### Connection with the Accumulated Disturbance Problem

Returning now to the problem formulated in the first paragraph, as far as the approximation to the final section of the trajectory is concerned, our problem is directly related to that of Bulgakov<sup>10</sup> on the accumulation of disturbances in a dynamic system.

Introducing the substitution  $z = x - \psi(t)$  into the system of eqn (1) we transform it into the form

$$\frac{dz}{dt} = Z(z, \eta, t) + r(c, y, \psi(t), \eta(t), t) \quad (8)$$

where

$$Z(z, \eta, t) = f(z + \psi(t), \eta(t), t) - f(\psi(t), \eta(t), t)$$

$$r(c, y, \psi(t), \eta(t), t) = f(\psi(t), \eta(t), t) - \psi'(t) + u(c, y, t) = r(t)$$

System (8) is a system of equations for perturbed motion, the function  $r(t)$  determines according to eqn (2) the approximation error of the programming or control functions, and the deviation of the solution  $z(t)$  of system (8) from zero coincides with the deviation of the solution  $x(t)$  of system (1) from the given function  $\psi(t)$ .

If system (8) is linear, then for  $z(0) = 0$ ,  $0 < t \leq T < \infty$  we have  $z(t) = Ar(t)$ , where A is a linear-bounded operator transforming the function  $r(t)$  into the functions  $z(t)$ . If  $\|A\|$  is the norm of the operator, then  $\|z(t)\| \leq \|A\| \|r(t)\|$  is obtained. The latter relation is also the most general expression of the solution to the problem of disturbance accumulation. By taking various norms for  $r(t)$  and  $z(t)$  and computing  $\|A\|$ , the actual inequalities that solve this problem<sup>11, 12</sup> are obtained.

### Connection with the Theory of Approximations

If as an optimum criterion that of the minimum error  $r(t)$  (in any dimension) is taken, then the problem of realizing the given trajectory reduces to a problem in the theory of approximations. This problem is most effectively solved in the case where  $r(t)$  depends linearly on the programming parameters and functions, and where we require a minimum of the mean-square approximation error. In this case the elementary rules of the theory of mean-square approximations are used for computing the control. It should be observed that here two essentially different cases are met. In the first case, by selecting the programming parameters from a sufficiently large number of the latter the approximation error can be made as small as required, i.e. the given motion can be achieved as accurately as necessary. In the second case the error of approximation cannot be made less than a certain value. Here it is worth while to state the problem of simultaneously choosing optimal values for the parameters and optimal programming functions. The success of such a choice depends, roughly speaking, on how well the given trajectory fits into linear subspaces in the various dimensions<sup>11</sup>.

### Trajectory Realization and Stability Theory

If it is wished to approximately realize motion along a given trajectory for the whole interval  $0 < t < \infty$ , certain difficulties

arise. It can readily be seen that such an approximate realization is possible if the zero solution of the system

$$\frac{dz}{dt} = Z(z, \eta, t) \quad (9)$$

is stable in relation to continuously acting disturbances that are limited relatively to the dimension in which the approximation error  $r(t)$  is evaluated. There exist<sup>13</sup> stability criteria related to continuously acting disturbances limited in modulus or in mean value. Stability criteria can easily be deduced<sup>14</sup> for use with continuously acting disturbances limited in their mean square, which are of most interest in our problem. However, in solving the problem it was required to find convenient evaluations of continuously acting perturbations that were simultaneously evaluations of approximation errors.

Such evaluations<sup>14</sup> were found and it turned out to be best to make them in the dimension of space  $M$  with norm

$$\|r(t)\|^2 = \sup_{0 < k < \infty} \int_{kT}^{(k+1)T} |r(t)|^2 dt$$

where  $|r(t)|$  denotes the length of the vector  $r(t)$ . Massera was the first<sup>15</sup> to point out the important role of the space  $M$  in stability theory.

Dwelling further on a question related to stability theory, the operating mode  $\psi(t)$  is called stable in relation to the system  $\dot{x} = X(x, t)$  if the zero solution of the system

$$\dot{z} = X(z + \psi(t), t) - X(\psi(t), t)$$

is asymptotically stable. From the preceding argument it is clear that only stable operating modes can claim to give a good approximation. Unfortunately few criteria for operating mode stability have so far been derived in relation to this system. Clearly if the basic system is linear and asymptotically stable, then any operating mode will be stable relative to it. The same property is possessed by the systems considered by Krasovskiy in his paper<sup>16</sup> (theorem 3.1). These systems are determined by the fact that for each of them a constant symmetrical matrix  $A$  can be defined having positive eigenvalues such that the symmetrized matrix

$$[B_{ik}] = \left[ \left( A \frac{\partial X}{\partial x} \right)_{ik} + \left( A \frac{\partial X}{\partial x} \right)_{ki} \right], \left( \frac{\partial X}{\partial x} \right)_{ik} = \frac{\partial X_i}{\partial x_k}$$

has negative eigenvalues  $\mu_i$  satisfying the inequality  $\mu_i < -d$ , where  $d > 0$  at all points of the space  $-\infty < x_i < \infty, 0 \leq t < \infty$ .

The interesting result obtained by Letov<sup>17</sup> is also noted, concerning non-linear control systems with parameters that vary only slightly. He has proved for a large class of systems of great importance in control engineering that the stability of a given operating mode implies the stability of all sufficiently close modes. In this case the closeness of the modes is assessed by the magnitude of the modulus of the difference between the programming functions.

Probably further results in this direction can be obtained on the basis of both existing and new criteria for asymptotic stability of linear systems with variable coefficients. It can easily be verified that in the unidimensional case Krasovskiy's criterion is a necessary and sufficient condition for the stability of any mode. It would be interesting to know to what extent this criterion is necessary for systems of a higher order.

## Realization of Periodic Motions

Now let the right-hand side of the system (1) and also the function  $\psi(t)$  be periodic in  $t$  with period  $T$ . Assuming that the zero solution of system (9) is asymptotically stable to a first approximation, we can again formulate the conditions for a given motion to be realizable with the required accuracy. But in this case these conditions can be set more simply, since here the dimension in space  $M$  is given by

$$\|r(t)\|^2 = \int_0^T |r(t)|^2 dt$$

Furthermore it can be shown that even in the presence of an approximation error different from zero there exists an asymptotically stable periodic motion lying within an  $\varepsilon$  neighbourhood of the given periodic motion.

It should be observed that the results obtained can be extended without difficulty to the case where the motion to be realized is discontinuous, or more accurately has discontinuities of the first sort<sup>14</sup>. In this case the programming functions will appear as the sums of ordinary functions and linear combinations of  $\delta$  functions.

## Programme Control of Random Processes

Up to now attention has not been directed to the external influence or, more precisely, disturbance  $\eta(t)$ . Normally  $\eta(t)$  is a random function, and so the actual mode of operation will be a random process. Naturally in this event the programmed mode is also random. The extension of the preceding results to the case of stochastic differential equations presents no difficulties, provided the following points are borne in mind. A random quantity, as is known, may be determined as a measurable function defined in some choice space  $\Omega$  (or space of elementary events). It is easy to see that the space  $\Omega$  can be constructed in such a way that it is the choice space for all random functions  $\eta(t)$ ,  $\xi(t)$  and  $x(t)$  occurring in the equation

$$\frac{dx}{dt} = f(x, t, \eta(t)) + u(t, \xi(t)) \quad (10)$$

where  $\xi(t)$  is the distortion of the disturbance  $\eta(t)$  (see Figure 1).

If a norm is defined by any means in the linear space of random quantities (as in the space of measurable functions defined in the choice space  $\Omega$ ), then differential eqn (10) is transformed into a differential equation given in the linear normalized space  $R$ , whose elements are random vectors. Here one should take as initial vectors in the solution of Cauchy's problem not only deterministic vectors but also any other random vectors from  $R$ , while the derivative and integral of a random function *w.r.t. t* should be understood as the derivative and integral in Bochner's sense. In particular, if the mathematical expectation of the square of the length of the vector is taken as the square of the norm of a random vector, then the concept of the derivative and integral of a random function coincides with the generally accepted one.

It should be observed that the theory of differential equations in a Banach space is well developed at the present day. By making use of this theory, one can readily formulate conditions for the existence, uniqueness and extensibility of solutions<sup>18</sup>, and consider questions of stability<sup>19</sup> or questions of the ex-

stence and research of periodic motions<sup>20</sup>. All this enables the setting up of a completely analogous statement of the problem of realizing random processes and to obtain results identical to those presented above<sup>21</sup>.

The reduction or elimination of the effect of disturbance by its continuous tracking has found wide application in the theory of automatic control, mainly in the theory of composite control systems. This theory uses the so-called invariance principle developed by Academicians Luzin and Kulebakin, which has served as the starting point for a large number of papers on automatic control theory that have important applications.

### Realization of Processes by means of Systems with Many-valued Characteristics

Barbashin and Alimov<sup>22</sup> have shown how to reduce systems of differential equations with relay-type hysteresis, and in general many-valued characteristics to a differential equation in a normalized linear space. Thus in this case also all the preceding results can be obtained by the same method as was indicated for the programming of random processes.

### Conclusions

It has been seen in this paper that the accuracy of approximation to the trajectory depends on the degree of stability of the zero solution of the system (10). The better this stability, as judged according to any of the existing quality criteria, the smaller effect will approximation errors have on the deviation of the trajectory from the given one. Thus the problem of improving the response of programme control turns on the problem of increasing the stability of motion. Here, in particular, the theory of programme control again comes into contact with the theory of optimal control.

### References

- BELLMAN, R. Notes on control processes, Pt I. On the minimum of maximum deviation. *Quart. appl. Math.* 14 (1957)
- PONTRYAGIN, L. S., BOLTYANSKIY, V. G., GAMKRELIDZE, R. V., and MISHCHENKO, YE. F. *The Mathematical Theory of Optimal Processes*. Fizmatizdat (1961)
- ROZONER, L. I. Pontryagin's maximum principle in the theory of optimal systems, Pt II. *Automat. Telemekh.* 20, No. 11 (1959)
- LETOV, A. M. The analytical design of controllers, Pt I. *Automat. Telemekh.* 21, No. 4 (1960)
- LETOV, A. M. The analytical design of controllers, Pt II. *Automat. Telemekh.* 22, No. 4 (1961)
- BARBASHIN, YE. A. On the approximate realization of motion along a given trajectory. *Automat. Telemekh.* 22, No. 6 (1961)
- ROYTENBERG, YE. N. Some problems in the theory of dynamic programming. *Prikl. Matem. Mekh.* 23, No. 4 (1959)
- BARBASHIN, YE. A. On a problem in the theory of dynamic programming. *Prikl. Matem. Mekh.* 24, No. 6 (1960)
- YEMEL'YANOV, S. V., and FEDOTOVA, A. I. The design of optimal automatic control systems of the second order using limiting values of the elements of the control circuit. *Automat. Telemekh.* 21, No. 12 (1960)
- BULGAKOV, B. V. On the accumulation of perturbations in linear oscillatory systems with constant parameters. *Dokl. Ak. Nauk SSSR* 51, No. 5 (1946)
- BARBASHIN, YE. A. The evaluation of the mean-square deviation from a given trajectory. *Automat. Telemekh.* 21, No. 7 (1960)
- BARBASHIN, YE. A. The evaluation of the maximum of the deviation from a given trajectory. *Automat. Telemekh.* 21, No. 10 (1960)
- GERMAIDZE, V. YE., and KRASOVSKIY, N. N. On stability in the presence of continuously-acting perturbations. *Prikl. Matem. Mekh.* 21, No. 6 (1957)
- BARBASHIN, YE. A. On the construction of periodic motions. *Prikl. Matem. Mekh.* 15, No. 2 (1961)
- MASSERA, J. L., and SCHÄFFER, J. J. Linear differential equations and functional analysis, Pt I. *Ann. Math.* 57, No. 3 (1958)
- KRASOVSKIY, N. N. Stability with large initial perturbations. *Prikl. Matem. Mekh.* 21, No. 3 (1957)
- LETOV, A. M. *The stability of non-linear controlled systems*. 1955. GITTL
- KRASNOSELSKIY, M. A., and KREIN, S. G. Non-local existence theorems and uniqueness theorems for systems of ordinary differential equations. *Dokl. Akad. Nauk SSSR* 102, No. 1 (1955)
- MASSERA, J. L. Contributions to stability theory. *Ann. Math.* 64, No. 1 (1956)
- MASSERA, J. L., and SCHÄFFER, J. J. Linear differential equations and functional analysis, Pt II. Equations with periodic coefficients. *Ann. Math.* 69, No. 1 (1959)
- BARBASHIN, YE. A. Programme control of systems with random parameters. *Prikl. Matem. Mekh.* 25, No. 5 (1961)
- BARBASHIN, YE. A., and ALIMOV, YU. I. Contribution to the theory of dynamic systems with non-single-valued and discontinuous characteristics. *Dokl. akad. Nauk SSSR* 140, No. 1 (1961)
- BARBASHIN, YE. A., and ALIMOV, YU. I. Contribution to the theory of relay-type differential equations. *Izv. Vyssh. Ucheb. Zav. Matemat.*, No. 1 (26), (1962)
- MILSTEIN, G. N. On the approximate realization processes. *Prikl. Matem. Mekh.* 26, No. 4 (1962)

### DISCUSSION

I. FLÜGGE-LOTZ, *Engineering Mechanics Division, Stanford University, Stanford, California, U.S.A.*

The author mentions the work by Yemelyanov and Fedotova, which promises a good follow-up of slowly timed varying inputs. The switching law is non-linear. However, since  $\text{sgn}[\varepsilon(Tp + 1)\varepsilon] = \text{sgn} \varepsilon \text{sgn}(Tp + 1)$  it can be realized easily. In 1957 Charles F. Taylor and I developed a control system which allows a good follow-up of inputs with strong slope discontinuities (sawtooth curves). The system is given by the following equations where  $f(t)$  is a given input.

$$y'' + 2\zeta(1 + \beta_m)y' + (1 + \gamma_n)y = f(t)$$

with 
$$\beta_m = \beta \text{sgn}(y'e) - \beta \text{sgn}(y'e')$$
$$\gamma_n = -\gamma \text{sgn}(ye) - \gamma \text{sgn}(ye')$$

Extensive testing of that system is reported in *NACA Tech. Rep. 1391* (1958). Since Yemelyanov's and Fedotova's switching law is so attractively simple, I would be interested in knowing whether experiments have been conducted to show the follow-up and its limitations.

As to third-order systems, I assume that the difficulty lies in a reasonably accurate subdivision of the three-dimensional space in

regions in which the system is governed by linear differential equations with desirable parameters. Did the authors consider using digital control?

YE. A. BARBASHIN, *in reply*

The example Professor Flügge-Lotz has reported shows that the optimum system with respect to the zero input may prove useless for an uneven action. This fact is well known by experts of the theory of automatic control. Examples like that might obviously be given in great numbers. In order to give an example, I resorted to the paper by Yemelyanov and Fedotova, which presents the optimum solution of the problem.

This paper does not deal with systems of variable structure. These are discussed in the paper by Petrov, Ulanov and Yemelyanov. It seems therefore to be more appropriate to refer to Mr. Yemelyanov who is investigating the system of variable structure. Since the field of mathematics deals also with these systems, I refer to the papers published in the journal *Automatika i Telemekhanika* 10, (1963) Nos. 1, 5, 7. In Nos. 1 and 7 will be found the results of the experiments concerning the method suggested. Yemelyanov's paper clearly shows the advantage of similar systems.

As to the application of discrete computing devices for the system of variable structure, this problem has been theoretically developed and is now in the stage of experimentation.



# The Realization of Optimal Programmes in Control Systems

G. S. POSPELOV

## Summary

Optimum control problems are, as a rule, *a priori* programmes which are formed either by mathematical programming methods (maximum principle, dynamic programming, etc.), or by control theory methods. A mathematical model is required, but since this model does not coincide with the actual process, additional devices and feedbacks are required for optimization. Several examples are given which illustrate how optimum programmes can be formed.

## Sommaire

Les systèmes d'optimisation sont en général des programmes *a priori* dérivés, soit de méthodes mathématiques (principe du maximum, programmation dynamique, etc.), soit de méthodes déduites de la théorie de commande. Un modèle mathématique du système est nécessaire. Cependant puisque ce modèle ne coïncide pas avec la réalité du processus à optimiser, des contre-réactions et des dispositifs supplémentaires sont nécessaires pour l'optimisation. Différents exemples sont donnés qui illustrent comment des programmes d'optimisation peuvent être établis.

## Zusammenfassung

Probleme der optimalen Regelung führen im allgemeinen zu „*a priori*-Programmen“, die entweder durch mathematische Programmiermethoden (Maximumprinzip, dynamische Programmierung usw.) oder durch Verfahren der Regelungstheorie gebildet werden. Man braucht ein mathematisches Modell, aber da dieses Modell nicht mit der wirklichen Strecke übereinstimmt, werden zur Optimierung weitere Geräte und Rückführungen benötigt. Für die Aufstellung von Optimalprogrammen werden einige Beispiele angegeben.

Methods of mathematical programming [the term is used to mean the application of mathematics to the practical activity of planning, development, decision-making etc., and is a natural generalization of such concepts as linear (or non-linear) dynamic programming] are spreading to all branches of the national economy, engineering, industry, agriculture and so on. This presupposes the development of mathematical models of the events or sets of controlled plants which require to be controlled. Once the aim of control has been formulated, the task is to determine the optimum strategy of control or a programme of effects upon the controlled plants produces in some sense the optimal result.

It must be emphasized that the programming methods determine the strategy of control or a *a priori* programme. The degree of coincidence between the actual result or process produced by control and the result or process anticipated from the *a priori* programme, is indicative, in particular, of the perfection of the mathematical model or of our knowledge of the controlled plant.

However, a mathematical model is a model and not the phenomenon itself, and, apart from this, during the process of

realizing the *a priori* programme, the controlled plant can be affected by a variety of factors and perturbations which are not taken into account in the model. This can lead to deviations, and sometimes to considerable deviations, from the programme results, which by definition are optimal.

If the programme is time scheduled, use can be made of feedback to correct the effect of perturbations and inaccuracies in the mathematical description so as to ensure an actual programme closer to the optimal one.

Those most completely represented by mathematical models are control systems. Taking their case as an example, we will consider the possible ways of realizing optimal programmes, in this instance, control programmes.

A mathematical model of a control system is usually formed by means of ordinary differential equations. The control programme is broadly defined to cover the planning of the dynamic characteristics of the control system, its programme of operation, etc. In all cases it is assumed that the system is provided with complementary feedbacks which improve the realization of the predetermined programme or *a priori* programme.

(1) The desired dynamic characteristics of a system are realized by complementary self-adjusting circuits, which in this case are complementary feedbacks which improve the realization of the predetermined programme of the control system of operation. Figure 1 shows the well-known self-adjusting system of an automatic pilot which controls the angle of pitch of an aircraft<sup>1</sup>. The self-adjustment circuit changes the gain of the angular velocity circuit so that the margin of stability of this circuit is maintained constant. The correcting circuit 2 is selected to obtain a sufficiently high gain  $K$ . Under these conditions the transfer function of the closed angular velocity circuit is close to unity. Therefore, despite the variation of the properties of the controlled plant (owing to changes in flying conditions), the

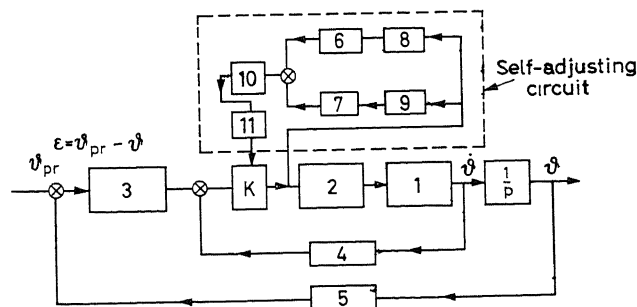


Figure 1.  $\theta$  - angle of pitch;  $\theta_{pr}$  - programme value of pitch angle; 1 - controlled plant; 2 - correcting circuit; 3 - model; 4, 5 - measuring devices for the angular velocity  $\dot{\theta}$  and the angle  $\theta$ ; 6, 7 - detectors; 8, 9 - high and low pass filters; 10 - servo motor; 11 - limiter

dynamic properties of the angle of pitch circuit will be determined by the transfer function of the model, i.e. in all cases they will be quite close to the predetermined or planned properties. Another example is the self-adjusting control system with extremal tuning of the correcting circuits<sup>2</sup>. Both examples refer to continuously operating control systems.

A somewhat special problem arises in the preservation of planned dynamic properties for 'single-action' systems<sup>3</sup> for which the behaviour is significant on a finite interval  $t$  ( $0 \leq t \leq t_1$ ), and for which the operating process is, as a rule, a transient process. Here one meets with the problem of maintaining a desired nature of transient behaviour, or a programme of motion of the representative point in phase space, on condition that the mathematical model does not exactly describe the dynamic properties of the controlled plant, nor the perturbations acting on the latter during the motion. Several possible ways of solving this problem are now indicated with simple examples.

Let the mathematical model of the controlled plant be represented in the form

$$\dot{x} = u \quad (1)$$

where  $x$  is the output coordinate and  $u$  is the controlling action.

Given the equation of the controller under the form

$$u = -a_0 x \quad (2)$$

the equation of the mathematical model of the system as a whole is

$$\dot{x} + a_0 x = 0 \quad (3)$$

According to (3), for any initial condition  $x_0$ , the process of motion is characterized by an exponential with the exponent  $-a$ . Now suppose that there is a suspicion that, in fact, the control object is described by the equation

$$\dot{x} = f(x, u, t) \quad (4)$$

where

$$f(x, u, t) = -a(t) \cdot \phi(x) + F(t) + u \quad (5)$$

$a(t)$ ,  $F(t)$  are random functions and  $\phi(x)$  is also random. Here it is known beforehand that  $|a(t)\phi(x) + F(t)| < |u|$ . In this situation one can make a decision concerning the discrete control of the plant, such that at each step it is possible to control the fulfilment of the *a priori* programme, which is expressed as a function of time in the following manner:

$$x = x_0 e^{-at} \quad (6)$$

With discrete control we require for control a relationship between the value of  $x(t)$  and the value of this coordinate at the instant of time  $t + \Delta t$ , i.e. the quantity  $x(t + \Delta t)$ . According to (6) this programme relationship is given by the relation

$$x(t + \Delta t) = x(t) \cdot e^{-a\Delta t} \quad (7)$$

where  $\Delta t$  is the interval of discreteness or the step of control. Using an analogy between the numerical solution of differential equations by difference methods and the discrete control of controlled plants, one writes the equation (4) in discrete form

$$x(t + \Delta t) = x(t) + f\left(t + \frac{\Delta t}{2}\right) \cdot \Delta t \quad (8)$$

where

$$f\left(t + \frac{\Delta t}{2}\right) = f\left[x\left(t + \frac{\Delta t}{2}\right), u\left(t + \frac{\Delta t}{2}\right), t + \frac{\Delta t}{2}\right]$$

The discrete form (8) of the solution of eqn (4) is used in the method proposed by Bashkirov. (The method of Bashkirov is described in the monograph by Popov<sup>4</sup>.) According to eqn (8), by measuring the value of  $x(t)$  at each step one can select the increment  $\Delta u$  at the instant  $t + (\Delta t/2)$  such that  $x(t + \Delta t)$  is governed by condition (7). The discrete form (8) is convenient in that the interval  $\Delta t/2$  is given in the procedure for calculating  $\Delta u(t + \Delta t/2)$ . The information for calculating  $\Delta u(t + \Delta t/2)$ , apart from the known value of the desired  $x(t + \Delta t)$ , is obtained from the preceding values of  $\Delta u$  and  $x$ . In the general case  $\Delta u(t + \Delta t/2)$  is calculated by the formula:

$$\Delta u\left(t + \frac{\Delta t}{2}\right) = \Delta u\left(t - \frac{\Delta t}{2}\right) \cdot \psi\left[x(t + \Delta t), x\left(t - \frac{\Delta t}{2}\right), x(t - \Delta t)\right] \quad (9)$$

The form of the function  $\psi$  depends on the particular theory of extrapolation which is adopted.

The information about the preceding values of  $x$  and  $u$  also includes information about changes in the properties of the plant and of the perturbation  $F(t)$ . The use of this information for calculating  $\Delta u(t + \Delta t/2)$  represents the additional feedback signals, or self-adjusting signals, and makes it possible to realize more accurately the desired programme of motion<sup>6</sup>.

Equation (4) and its results can be generalized without difficulty to multi-dimensional systems of any order. In this case the equation of the controlled plant in the vector form is

$$\frac{dX}{dt} = f(X, U, t) \quad (10)$$

where  $X$  is the vector with the components  $x_i$  ( $i = 1, 2, \dots, n$ ),  $f$  is the vector with the components  $f_i$  ( $i = 1, 2, \dots, n$ ), and  $U$  is the control vector with the components  $u_i$  ( $i = 1, 2, \dots, \gamma$ );  $\gamma \leq n$ .

The achievement of planned dynamic properties of single-action systems can also be realized by a continuous control. Suppose, for example, that the mathematical model of the controlled plant is written in the form

$$\ddot{x} + a_1 \dot{x} = u \quad (11)$$

and  $|u| \leq u_0$ .

Suppose also that it is required to realize the system with maximum operating speed. According to Pontryagin's maximum principle<sup>5</sup>, the equation of the controller is of the form

$$u = -u_m \operatorname{sign}[x + f(a_1, \dot{x})] \quad (12)$$

However, there is a suspicion that in fact the controlled plant can be described by the equation

$$x + a_1^*(t) \cdot \dot{x} + a_0^*(t) \cdot x = u + F(t) \quad (13)$$

In view of the incomplete information about  $a_1^*(t)$ ,  $a_0^*(t)$  and  $F(t)$  it is impossible to accept the control law of type (12) which would ensure the maximum operating speed.

In view of this one proceeds as follows, forming the acceleration control circuit  $\ddot{x} = n$  by means of the controlling action  $u$  (Figure 2). If the pass band of this circuit is sufficiently high the error  $\varepsilon_n = n_{pr} - n$  will be close to zero and the programme acceleration will be equal to the actual acceleration. In more complex cases the acceleration control circuit, like the pitch angle control circuit (Figure 1), can be a self-adjusting circuit. If now the programme acceleration is close to the actual accelera-

tion, any desired variation of the coordinate  $x$  and its derivative may be required. Thus, to form the system of maximum operating speed in accordance with the mathematical model (11), it is sufficient to put

$$\ddot{x}_{pr} \cong \ddot{x} = -a_1 \dot{x} - u_1 \operatorname{sign}[x + f(a_1, \dot{x})] \quad (14)$$

The block diagram which realizes (14) is shown in Figure 3. In expression (14)  $u_1$  is always less than  $u_0$  since some part of the control resource  $u_0 - u_1$  goes to compensate the perturbation

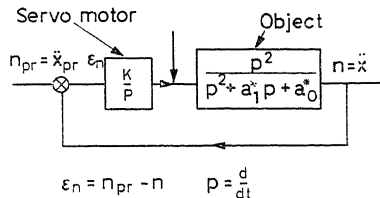


Figure 2

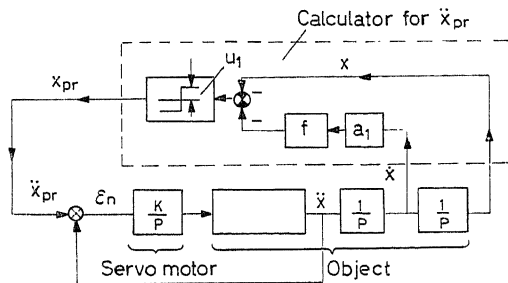


Figure 3

$F(t)$  and to compensate the difference between the coefficients  $a_1^*(t)$  and  $a_0^*(t)$  on the one hand and the coefficients of the mathematical model  $a_1$  and  $a_0 = 0$  on the other. Thus, at the expense of some reduction of operating speed (since  $u_1 < u_0$ ) a definite realization of the programme for the optimum transient process is obtained.

Any other law of variation of the coordinate  $x$  can be required in this example. It may, for example, be required that the transient process should take place in accordance with the solution of a linear equation with constant coefficients

$$\ddot{x} + a_1 \dot{x} + a_0 x = 0 \quad (15)$$

For this, it is obviously necessary to put  $\ddot{x}_{pr} = -a_1 \dot{x} - a_0 x$ .

Figure 4 shows oscillograms which have been obtained on the electronic simulator for the case when  $|a_0^*(t)| \leq 0.05$ ;  $|a_1^*(t)| \leq 1.0$ ;  $a_1 = 0.4$ ;  $a_0 = 0.04$ . The gain of the servo motor of the acceleration control circuit was taken as 10 l/sec. It will be seen from the oscillogram that the perturbation  $F(t)$  and the fluctuations of the coefficients  $a_1^*(t)$  and  $a_0^*(t)$  have no effect on the course of the coordinate  $x$  which is governed by the solution of eqn (15).

The results explained by this example are also capable of very wide generalization. The generalization consists in that for a known indeterminacy of the properties of the controlled plant and of the acting perturbations it is advisable to organize a self-adjusting subsystem of rapidly varying coordinates of the controlled plant or of its higher-order derivatives. After the programme variation of the rapidly varying coordinates or of their higher-order derivatives has been largely determined by this subsystem, the law governing the variation of the slowly varying coordinates or lower-order derivatives of the output magnitude of the controlled plant can be built as desired. The additional feedbacks which make it possible to realize the required programme of dynamic properties of the system are, in the example under consideration, the feedbacks amongst which are the self-adjusting circuits for acceleration control.

Very often the realization of desired dynamic properties for single-acting systems is handicapped by unfavourable combinations of initial conditions. In non-linear systems these unfavourable combinations of initial conditions can lead to instability of the process for a given realization. The effect of unfavourable combinations of initial conditions can be eliminated by changing the initial values of the coordinates and by the formation of special signals which act on the system and which are functions of the initial conditions. Briefly, this means creating special feedbacks with respect to the initial conditions. The idea of using feedback with respect to the initial conditions has already been published in a paper by the author<sup>6</sup>.

(2) In developing systems with programme control of the output coordinates of the controlled plant use may, to a large extent, be made of the foregoing ideas and methods which relate to the realization of programmed dynamic properties of control systems.

Suppose, for example, that it is required to vary according to the programme  $g_{pr}(t)$  the output coordinate  $x(t)$  of the controlled plant (Figure 5). For this the input of a closed system consisting of the controlled plant and the controller receives the programme signal  $g_{pr}(t)$ . For a system with a high pass band, if no perturbations are present, it is well known that  $x \cong g_{pr}(t)$ .

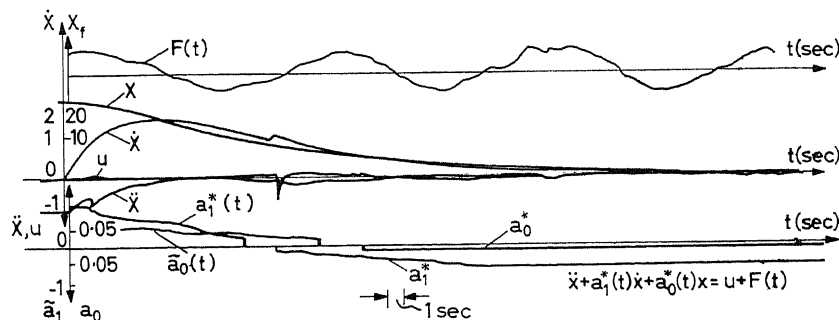


Figure 4

However, a random perturbation which is not taken into account can considerably distort the desired programmed variation of  $g_{pr}(t)$ . In order to fulfil more accurately the programme, an additional feedback is formed (shown by the dotted line in Figure 5) and the programme correction circuit *abcdega* is thereby formed. The programme signal  $g_{pr}(t)$  is compared with the

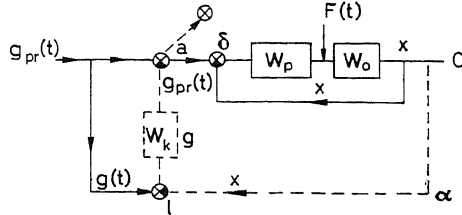


Figure 5.  $W_p$  - controller;  $W_0$  - plant;  $W_k$  - self-adjusting correction circuit with a high gain

actual signal and the difference signal acts at the input to the fundamental system *via* a self-adjusting correction circuit with a high gain  $W_k$ . The correction circuit may consist of the elements 2, K, 6, 7, 8, 9, 10 and 11 which are shown in Figure 1. Assuming, for the sake of simplicity,  $W_k = K$ , the following operator relationship is obtained between the input and output for the circuit of Figure 5:

$$x = \frac{\phi(p)(1+K)}{1+K\phi(p)} g_{pr}(t) + \frac{\phi_f(p)}{1+K\phi(p)} F(t) \quad (16)$$

where

$$\phi(p) = \frac{W_p W_0}{1 + W_p W_0} \quad \text{and} \quad \phi_f(p) = \frac{W_0}{1 + W_p W_0}$$

and expression (16) can be written as

$$x = \frac{\phi(p) \left( \frac{1}{K} + 1 \right)}{\frac{1}{K} + \phi(p)} g_{pr}(t) + \frac{\phi_f(p)}{\frac{1}{K} + \phi(p)} \frac{1}{K} F(t) \quad (17)$$

It will be seen from (17) that if  $K \rightarrow \infty$

$$x = g(t)_{pr} \quad (18)$$

independently of the action of the perturbation  $F(t)$  and the fluctuations of the parameters of the controlled plant. It is understood that in this case condition (18) is fulfilled approximately since  $K = \infty$  is not realizable in actual conditions.

Another example of programme control is the method of stabilizing acceleration (Figures 2 and 3) with subsequent construction of the desired programmed variation of the coordinate  $x_0$  by means of a computer.

Using this method the 'logarithmic navigation'<sup>7</sup> can be realized when the acceleration according to the programme  $k \ddot{x}/x$ , and consequently, the coordinate  $x$  is the solution of the differential equation

$$x \ddot{x} - k \dot{x} = 0$$

A very important case of programme control is that when it is important to maintain a functional relationship between one coordinate and another. For example, the optimum programme, as regards operating speed, for the altitude and speed of an

aircraft, as calculated, for instance, by the method of dynamic programming, is a programme in the coordinates  $H$  and  $V$ , i.e. it is given as a functional relationship  $H_{pr} = H_{pr}(V_{pr})$  (Figure 6), both the quantities  $H$  and  $V$  here being the output coordinates of an aircraft controlled by the altitude rudder (the thrust of the engine is usually maximum in this case). The

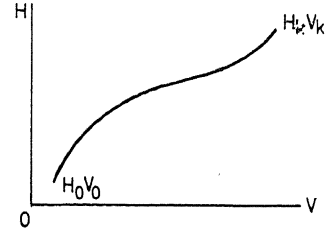


Figure 6

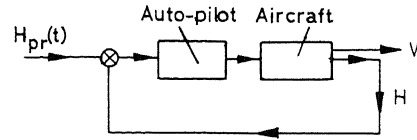


Figure 7

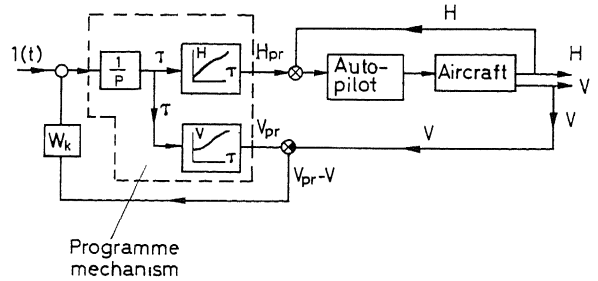


Figure 8

relationship  $H_{pr} = H_{pr}(V_{pr})$  can always be represented parametrically:

$$H_{pr} = H_{pr}(t)$$

$$V_{pr} = V_{pr}(t)$$

The altitude control circuit  $H$  can now be formed by the usual method (Figure 7). If the system is unaffected by perturbations and the calculated characteristics of the aircraft coincide with the actual characteristics, and if the atmosphere through which the aircraft is flying remains standard, the completion of the programme  $H_{pr}(t)$  will at the same time imply the completion of the programme  $V_{pr}(t)$ , and consequently of the programme relationship  $H_{pr} = H_{pr}(V_{pr})$ . However, if all the stated conditions are not fulfilled, the completion of  $H_{pr}(t)$  will not generally imply the fulfilment of  $V_{pr}(t)$ , and consequently the completion of  $H_{pr} = H_{pr}(V)$ . For the planned programme  $H_{pr} = H_{pr}(V)$  to be fulfilled with acceptable accuracy, it is necessary to introduce a programme correction circuit<sup>8</sup>. For this purpose the programme value of speed is compared with the actual speed and the difference in terms of the transfer function  $W_k$  changes the rate at which the programme is delivered, i.e. the speed of the clocks of the programme mechanisms  $H_{pr}$  and  $V_{pr}$  (Figure 8).

As a result the speed of the clock mechanism of the programme is not uniform and the programmes  $H_{pr}$  and  $V_{pr}$  become functions of some irregularly varying argument  $\tau$ , i.e.  $H_{pr}(\tau)$  and  $V_{pr}(\tau)$ . Elimination of the argument  $\tau$  again brings us back to the original relationship  $H_{pr}(V_{pr})$ . However, in so far as the rate of delivery of the programme signal  $H_{pr}$  at the input of the system conforms to the fulfilment of the speed programme, the accuracy of the realization of  $H_{pr} = H_{pr}(V_{pr})$  is substantially increased. A similar circuit can be constructed for the motion of some controlled plant along a prescribed unperturbed trajectory  $y_e = y_e(x_e)$  in the coordinates  $x, y$  (Figure 9). However, this report is confined to the plane problem. Suppose that the speed of the plant is  $V$  and that the orientation of the speed vector

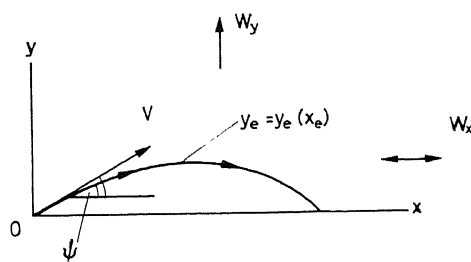


Figure 9

is characterized by the angle  $\psi$ . The obvious relationship between the coordinates  $x, y$  and the speed is expressed as follows

$$\dot{y} = V \sin \psi + W_y \quad (19)$$

$$\dot{x} = V \cos \psi + W_x \quad (20)$$

where  $W_y$  and  $W_x$  are perturbations in the form of speeds of displacement of the environment relative to the system of coordinates  $x, y$ . [In the formulae (19) and (20) the actual values of the coordinates of the controlled plant are used. The values of the desired unperturbed trajectory are denoted as  $x_e$  and  $y_e$ .] Consider the kinematic problem, i.e. suppose that the angle  $\psi$  of the speed vector can be arranged arbitrarily. On this assumption the control circuit for the coordinate  $y$  is formed. Here it is required that

$$\sin \psi = k_e \varepsilon \quad (21)$$

where

$$\varepsilon = y_{pr} - y \quad (22)$$

the term  $y_{pr} = y_{pr}(t)$  here being the programme value of the coordinate  $y$ , which does not coincide, as will be seen below, with the unperturbed value  $y_e = y_e(x_e)$ .

The equation for the coordinate  $y$  is found from the equations (19), (21) and (22):

$$\dot{y} + V k_e y = V k_e y_{pr}(t) + W_y \quad (23)$$

Assuming  $y = y_e + \Delta y$ , we obtain now the equation for the deviation  $\Delta y$  from the unperturbed motion.

$$\Delta \dot{y} + V k_e \Delta y = V k_e (y_{pr} - y_e) - \dot{y}_e + W_y \quad (24)$$

It is worthwhile to select the programme signal  $y_{pr}$  in accordance with the formula

$$y_{pr} = y_e + \frac{\dot{y}_e}{V k_e} \quad (24a)$$

For this value of the programme signal, eqn (24) becomes

$$\Delta \dot{y} + V k_e \Delta y = W_y \quad (25)$$

This implies that in the absence of the action  $W_y$  the deviation from the unperturbed trajectory will tend to zero. A constant action will cause a constant error.

It is obvious that a single control circuit according to the coordinate  $y$  cannot ensure the necessary control of the coordinate  $x$  or the fulfilment of the required programme  $y_e = y_e(x_e)$ . As the coordinate  $x$  varies according to the expression

$$x = V \int_0^t \cos \psi_e dt - V \int_0^t \tan \psi_e \cdot \Delta \dot{y} dt + \int_0^t W_x dt \quad (26)$$

The first term in eqn (26) is the desired unperturbed value of  $x = x_e$ , the second term can be limited, since it is determined by the error in the circuit for the stabilization of  $y$ , and the third term for  $W_x = \text{const}$  will continuously increase. In order to realize the programme of motion along the unperturbed trajectory it is necessary to proceed in the same way as in the previous case (see Figure 8), i.e. it is necessary to form, by measuring the error  $x_{pr} - x$ , a signal which acts on the speed of the programme mechanism  $y_{pr}(\tau)$  and  $x_{pr}(\tau)$ .

It should be noted that it is much simpler to correct the programme by varying the speed of the programme clocks if in the first example  $dV/dt > 0$ , and in the second example if  $dx/dt > 0$ . Generalizing, this method of correction to the programme of a system with  $n$  coordinates and  $r$  controlling devices, we shall note that in this case the argument of control (the non-decreasing coordinate  $V$  in the first example, and the non-decreasing coordinate  $x$  in the second) should be any constant sign form of system derivative<sup>9</sup>.

Frequently this form of coordinate originates naturally from the statement of the problem. For example, this is the case if it is required to control the ingredients of a mixture as a function of the volume of this mixture when this volume is varying in a monotonous way.

## References

1. MELLE, D. Application of adaptive flight control. *Symposium on Self-adjusting Systems*. Rome, April 1962
2. KRASOVSKI, A. A. The dynamics of continuous control systems with extremal self-adjustment of correcting devices. *Automatic and Remote Control*. 1961. London; Butterworths
3. POSPELOV, G. S. Concerning the principles of construction of various types of self-adjusting control systems. *Symposium on Self-adjusting Systems*, Rome. April 1962
4. POPOV, E. P. *Dynamics of Control Systems*. 1954. Moscow; GITTL
5. PONTRYAGIN, L. S., BOLTYANSKII, V. G., GAMKRELIDZE, R. V. and MISTCHENKO, E. F. *Mathematical Theory of Optimal Processes*. 1961. Moscow; Fizmatgiz
6. POSPELOV, G. S. Various methods of improving the quality of processes of regulation and control. *Symposium on Use of Computer Engineering in Automation of Production* (in Russian), 1961. Moscow; Mashgiz

- <sup>7</sup> Green, Logarithmic navigation for precise guidance of space vehicles. *IRE Trans.*, ANF-8, No. 2 (1961)
- <sup>8</sup> KOZIOROV, L. M., and KOROBKOV, M. N. A method of stabilizing the functional relationship between two interrelated variables by

means of one control device. *Iz. Akad. nauk U.S.S.R., OTN, Energetika i Avtomatika*, No. 4 (1961)

- <sup>9</sup> LETOV, A. M. *The Stability of Non-linear Control Systems*. 1955. Moscow; GITTL

## DISCUSSION

R. PETROVIC, *Institute for Automation and Telecommunications, Bircaninova 20, Belgrade, Yugoslavia*

It seems that the problem of deviations from the *a priori* optimal process dynamics due to small perturbations entering the process in some uncontrolled fashion, can be successfully investigated using the adjoint method and adjoint computational technique.

Small uncontrolled perturbations along the optimal state trajectory and/or along the optimal control trajectory can result either in a deviation from the desired process final state or in a change of the process performance index. To establish the dependence between the small perturbations along the optimal trajectories in  $X$  and  $U$  spaces, and the final state deviations or the functional changes, it is assumed that the real dynamics of the process are close to the *a priori* optimal dynamics. In this way the 'additional dynamics' along the optimal dynamics are the solution of a linear vector differential equation, in a matrix form

$$d/dt x_0(t), x(t), u(t) = A(t) x_0(t), x(t), u(t) \quad (1)$$

( $x_0(t)$  is the index of optimality as a function of time;  $[A(t)]$  is  $(1 + n + \gamma) \cdot (1 + n + \gamma)$  matrix of coefficients whose elements are the first partial derivatives of functions  $f_0, f_1 \dots f_n, \dot{u}_1 \dots \dot{u}_\gamma$  with respect to the state and control coordinates).

To the homogeneous linear vector differential equation (1) corresponds the adjoint vector differential equation

$$d/dt [\psi(t)] = -[A(t)]^T [\psi(t)] \quad (2)$$

Since the matrix product between the transposed adjoint solution and the solution of the original set is a scalar independent of time, one can obtain the relation between the solution families of the original set of eqn (1) at a fixed time  $t^* \in (t_0, t_1)$  and the solution of the adjoint set in time running backwards from the real time, started at  $t = t^*$ . Thus, in order to investigate the effects of all the small state and/or control perturbations on the deviation from the *a priori* final state it is sufficient to achieve the response of the adjoint system  $\psi(\tau)$ ,  $\tau = t_1 - t$ , by applying in a proper order the initial values to the components  $\psi_1 \dots \psi_n$ . Similarly, to analyse the sensitivity of the process performance index it is sufficient to find the response of the adjoint system  $\psi(\tau)$  by applying an initial value to the component  $\psi_0$ .

It should be pointed out that to find the analytical solutions of the adjoint set is generally impractical, but, from the computer mechanization point of view, it can be easily performed by means of models with time-varying coefficients<sup>1, 2</sup>.

## References

- <sup>1</sup> SCHMIDT, S. State space technique applied to the design of a space navigation system, *J.A.C.C.* (1962)
- <sup>2</sup> GAVRILOVIC, M., PETROVIC, R. and SILJAK, D. Adjoint method in the sensitivity analysis of optimal systems, *J. Franklin Inst.* No. 7 (July 1963)

# Some Bounds on Quantization Errors in Dynamic Programming Computations

J. J. G. GUIGNABODET

## Summary

The application of the Principle of Optimality leads very naturally to a computational procedure for the solution of optimal control problems. In the numerical treatment, however, as a result of the use of various interpolation schemes, errors are generated that propagate throughout the computation in a cumulative manner.

In this paper we formulate the basic computational problem and calculate upper bounds of these errors in the case where the state of the process under study is quantized and also in the case where, in addition, the search of the extremal is carried out over a finite set of discrete values of the control variables.

## Sommaire

L'application du principe d'optimalité conduit très naturellement à une méthode de calcul permettant de résoudre des problèmes de commande optimale. La solution numérique cependant, à cause de l'utilisation répétée de certaines formules d'interpolation est entachée d'erreurs qui se propagent tout en s'accumulant.

Dans cette note nous formulons le processus fondamental de calcul et nous déterminons la majoration des erreurs dans le cas où l'état du système étudié est quantifié et aussi dans le cas où, en plus, l'extremum est déterminé sur un ensemble fini de valeurs discrètes de la variable de commande.

## Zusammenfassung

Die Anwendung des Optimierungsprinzips führt ganz natürlich auf Berechnungsverfahren zur Lösung von Problemen der optimalen Regelungen. Die zahlenmäßige Behandlung ist allerdings wegen der Verwendung verschiedener Interpolationsmethoden mit Fehlern behaftet, die sich durch die ganze Rechnung in gesteigertem Maße fort-pflanzen.

In diesem Beitrag werden die grundlegenden Probleme der Berechnung formuliert und eine obere Fehlergrenze, sowohl für den Fall, daß der jeweilige Zustand des betrachteten Prozesses quantisiert wird, als auch für den Fall, daß außerdem der Extremwert durch eine begrenzte Anzahl diskreter Werte der Regelgröße gesucht wird, ausgerechnet.

A number of control problems may be reduced to finding an optimal sequence of decisions  $\{u_k\}$  which maximizes the functional

$$h_N(x) = \max_{\{u_k\}} \sum_{k=1}^N F(x_k, u_{k-1})$$

subject to constraint equations of the form

$$x_{n+1} = g(x_n, u_n)$$

$$u_k \in U$$

where  $g$  is the process transition function,  $x_n$  the state at the  $n$ th decision stage, and  $U$  a closed region in the control or decision space. Such a problem may be treated by the functional equation

technique of 'dynamic programming'<sup>1</sup>. However, because of the nature of the constraints, an analytical solution is generally not obtainable and one must resort to computational procedures<sup>2</sup>.

Consider the case where both  $x$  and  $u$  are scalars and assume that interest is in the values of  $h_N(x)$  for  $x$  in a closed interval  $X_N$ . Indeed,  $h_N(x)$  can be evaluated only at a finite subset of discrete values in that interval. Let  $X^*$  be the set of all even multiples of  $\lambda$ ,

$$X^* = \{0, \pm 2\lambda, \pm 4\lambda, \dots\} \text{ and define } X_N^* = X^* \cap X_N,$$

The functional equation formulation of the problem is now

$$h_n(x^*) = \max_u [F(g(x^*, u), u) + h_{n-1}(g(x^*, u))]$$

with the additional constraint  $x^* \in X^*$ .

It is clear that if  $h_N(x^*)$  is to be evaluated in this way for all  $x^* \in X_N^*$ , then  $h_{N-1}(x^*)$  must be evaluated for all  $x^* \in X_{N-1}^*$  where

$$X_{N-1}^* = X^* \cap X_{N-1}, X_{N-1} = [x_1^m, x_1^M]$$

$$x_1^m(x_1^M) = \min_{\substack{u \in U \\ x \in X_N}} (\max) g(x, u)$$

Thus,  $h_n(x^*)$  must be evaluated for all  $x^* \in X_n^*$  with

$$X_n^* = X^* \cap X_n, X_n = [x_{N-n}^m, x_{N-n}^M]$$

$$x_{N-n}^m(x_{N-n}^M) = \min_{\substack{u \in U \\ x \in X_{n+1}}} (\max) g(x, u)$$

The computation goes as follows:

$$h_1(x^*) = \max_u F(g(x^*, u), u)$$

is evaluated first at all  $x^* \in X_1^*$  and tabulated, then one proceeds with  $h_2(x^*)$ ,

$$h_2(x^*) = \max_u [F(g(x^*, u), u) + h_1(g(x^*, u))]$$

with  $x^* \in X_2^*$ . Since, in general,  $g(x^*, u) \notin X_1^*$  an approximate value  $h_2^e(x^*)$  is obtained by replacing  $h_1(g(x^*, u))$  by  $h_1(g(x^*, u) + l)$  in such a way that

$$(g(x^*, u) + l) \in X_1^*$$

$$l \in \Lambda, \quad \Lambda = (-\lambda, \lambda)$$

where the interval  $\Lambda$  is open on the left and closed on the right (or vice versa) to insure the uniqueness of  $l$  for all  $x^*$  and  $u$ .

This technique is then applied throughout the computational procedure by means of the equation

$$h_n^c(x^*) = \max_u [F(g(x^*, u), u) + h_{n-1}^c(g(x^*, u) + l)],$$

$$n = 2, 3, \dots, N$$

Let  $X = \cup_n X_n$ , it will be shown that if  $F(x, u)$  and  $g(x, u)$  satisfy Lipschitz conditions over  $X$

$$|F(x', u) - F(x'', u)| \leq K |x' - x''|^a$$

$$|g(x', u) - g(x'', u)| \leq K' |x' - x''|^b$$

$$x', x'' \in X, \quad u \in U$$

then

$$\lim_{\lambda \rightarrow 0} h_n^c(x^*) = h_n(x^*)$$

and an upper bound  $J_n(\lambda)$  of the error  $|h_n(x^*) - h_n^c(x^*)|$  will be calculated.

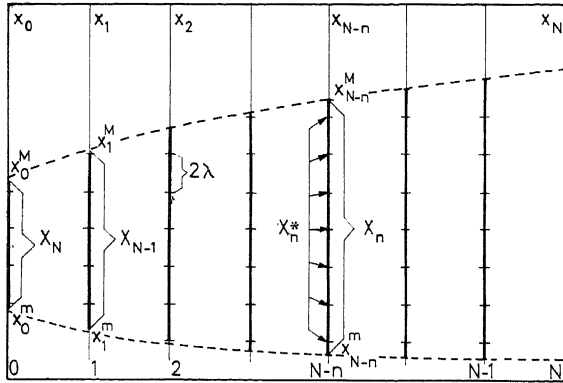


Figure 1. The sets  $X_n$  and  $X_n^*$

In order to get an upper bound as small as possible, the variations of  $F$  and  $g$  will also be bounded over each set  $X_n$  separately.

Namely,

$$|F(x', u) - F(x'', u)| \leq K_n |x' - x''|^a \quad (1)$$

$$|g(x', u) - g(x'', u)| \leq K'_n |x' - x''|^b$$

$$x', x'' \in X_n, \quad u \in U$$

Clearly,  $K_n \leq K$  and  $K'_n \leq K'$

$J_n(\lambda)$  is arrived at in the following way:

Since

$$\max_y |f(x, y) + h(x, y)| \leq \max_y |f(x, y)| + \max_y |h(x, y)|$$

and

$$|\max_y f(x, y) - \max_y h(x, y)| \leq \max_y |f(x, y) - h(x, y)|$$

Then, with the constraint  $(g(x^*, u) + l) \in X_{n-1}$ ,

$$\begin{aligned} & \max_{x^* \in X_n^*} |h_n(x^*) - h_n^c(x^*)| \\ & \leq \max_{\substack{u \in U \\ l \in \Lambda \\ x^* \in X_n^*}} |h_{n-1}(g(x^*, u)) - h_{n-1}(g(x^*, u) + l)| \\ & \quad + |h_{n-1}(g(x^*, u) + l) - h_{n-1}^c(g(x^*, u) + l)| \end{aligned}$$

$$\leq J_{n-1}(\lambda) + \max_{\substack{u \in U \\ l \in \Lambda \\ x^* \in X_n^*}} |h_{n-1}(g(x^*, u)) - h_{n-1}(g(x^*, u) + l)|$$

$$\leq J_{n-1}(\lambda) + J'_{n-1}(\lambda) = J_n(\lambda)$$

where  $J'_{n-1}(\lambda)$  is obtained in the following way, with  $(g(x^*, u) + l) \in X_{n-1}$

$$\begin{aligned} & \max_{\substack{u \in U \\ l \in \Lambda \\ x^* \in X_{n-1}^*}} |h_{n-1}(g(x^*, u)) - h_{n-1}(g(x^*, u) + l)| \\ & \leq \max_{\substack{u, u' \in U \\ l \in \Lambda \\ x^* \in X_n^*}} \{|F(g[g(x^*, u), u'], u') - F(g[g(x^*, u) + l, u'], u')| \\ & \quad + |h_{n-2}(g[g(x^*, u), u']) + h_{n-2}(g[g(x^*, u) + l, u'])|\} \\ & \leq J'_{n-1}(\lambda) \end{aligned}$$

Since

$$g[g(x^*, u), u'], \quad g[g(x^*, u) + l, u'] \in X_{n-2}$$

it follows that the upper bound of the first term is  $K_{n-2}(K'_{n-1}\lambda^b)^a$ . In the second term, the upper bound of the difference between the arguments of the functions  $h_{n-2}$  is  $K'_{n-1}\lambda^b$ ; therefore, one has the recursive relation

$$J'_{n-1}(\lambda) = J'_{n-2}(K'_{n-1}\lambda^b) + K_{n-2}(K'_{n-1}\lambda^{ab})$$

Now

$$h_1(x^*) = h_1^c(x^*) \quad \text{i.e.,} \quad J_1(\lambda) = 0$$

and, with  $x^* \in X_2^*$

$$\begin{aligned} & |F(g[g(x^*, u), u'], u') - F(g[g(x^*, u) + l, u'], u')| \\ & \leq K_0 K_1^a \lambda^{ab} = J'_1(\lambda) \end{aligned}$$

The result can thus be formulated as

**Theorem I**

If the functions  $F$  and  $g$  satisfy the Lipschitz conditions (1) and if the interval of quantization of the state variable  $x$  is  $2\lambda$ , then an upper bound  $J_n(\lambda)$  of the error  $|h_n(x^*) - h_n^c(x^*)|$  is

$$J_n(\lambda) = J_{n-1}(\lambda) + J'_{n-1}(\lambda), \quad n \geq 2$$

$$J'_{n-1}(\lambda) = J'_{n-2}(K'_{n-1}\lambda^b) + K_{n-2}(K'_{n-1}\lambda^{ab}), \quad n \geq 3$$

$$J_1(\lambda) = 0, \quad J'_1(\lambda) = K_0 K_1^a \lambda^{ab}$$

Explicitly,

$$\begin{aligned} J_n(\lambda) &= \lambda^{ab} (K_0 K_1^a + K_1 K_2^a + \dots + K_{n-2} K_{n-1}^a) \\ & \quad + \lambda^{ab^2} (K_0 K_1^a K_2^{ab} + \dots + K_{n-3} K_{n-2}^a K_{n-1}^{ab}) \\ & \quad + \lambda^{ab^3} (K_0 K_1^a K_2^{ab} K_3^{ab^2} + \dots + K_{n-4} K_{n-3}^a K_{n-2}^{ab} K_{n-1}^{ab^2}) \\ & \quad \vdots \\ & \quad + \lambda^{ab^{n-1}} K_0 K_1^a K_2^{ab} \dots K_{n-1}^{ab^{n-2}}, \quad n \geq 2 \end{aligned}$$

Clearly,

$$\begin{aligned} J_n(\lambda) &\leq K \{(n-1) \lambda^{ab} K'^a + (n-2) \lambda^{ab^2} K'^{a(1+b)} + \dots \\ & \quad + \lambda^{ab^{n-1}} K'^{a(1+b+b^2+\dots+b^{n-2})}\} \end{aligned}$$



If  $a = b = 1$ , this later relation becomes simply

$$\lambda K [(n-1)K' + (n-2)K'^2 + \dots + K'^{n-1}]$$

Since  $a$ ,  $b$ , and  $K$  are positive, it is seen that  $J_n(\lambda)$  goes to zero uniformly with  $\lambda$ .

In general, however, both  $x$  and  $u$  are quantized. Let  $U^*$  be the intersection of  $U$  and the set of all even multiples of  $\psi$ . The search of a maximum over the continuous range  $U$  is now replaced by the search of a maximum over a finite set  $U^*$ . One can write

$$h_n(x^*) = \max_{u^* + \rho} [F(g(x^*, u^* + \rho), u^* + \rho) + h_{n-1}(g(x^*, u^* + \rho))]$$

where  $u^* \in U^*$  and  $\rho \in \Psi$ ,  $\Psi = (-\psi, \psi)$  i.e., the search over the continuous range  $U$  is done by trying out successive discrete values  $u^*$  and letting  $\rho$  vary continuously over the range  $\Psi$ . Now

$$h_n^c(x^*) = \max_{u^*} [F(g(x^*, u^*), u^*) + h_{n-1}^c(g(x^*, u^*) + l)]$$

and

$$|F(x', u') - F(x'', u'')| \leq K_n |x' - x''|^a + K_n'' |u' - u''|^c \quad (2)$$

$$|g(x, u') - g(x, u'')| \leq K_n''' |u' - u''|^d$$

$$x', x'' \in X_n, \quad u', u'' \in U$$

Following the same approach as before, an upper bound  $J_n(\lambda, \psi)$  of the error  $|h_n(x^*) - h_n^c(x^*)|$  is derived. Setting  $\mu_n = K_n''' \psi^d + \lambda$ , the result is

### Theorem II

If the functions  $F$  and  $g$  satisfy Lipschitz conditions (1) and (2) and if the intervals of quantization of the state variable  $x$  and the control variable  $u$  are respectively  $2\lambda$  and  $2\psi$ , an upper bound  $J_n(\lambda, \psi)$  of the error  $|h_n(x^*) - h_n^c(x^*)|$  is

$$J_n(\lambda, \psi) = J_{n-1}(\lambda, \psi) + J'_{n-1}(\mu_n) + K''_{n-1} \psi^c + K_{n-1} K_n''' \psi^{ad},$$

$$J'_{n-1}(\mu_n) = J'_{n-2}(K'_{n-1} \mu_n^b) + K_{n-2} K'_{n-1} \mu_n^{ab}, \quad n \geq 3$$

$$J_1(\lambda, \psi) = K_0'' \psi^c + K_0 K_1''' \psi^{ab}, \quad J'_1(\mu_2) = K_0 K_1' \mu_2^{ab}$$

Explicitly,

$$J_n(\lambda, \psi) = \psi^c (K_0'' + K_1'' + \dots + K_{n-1}'')$$

$$+ \psi^{ad} (K_0 K_1''' + K_1 K_2''' + \dots + K_{n-1} K_n''')$$

$$+ K_0 K_1' \mu_2^{ab} + K_1 K_2' \mu_3^{ab} + \dots + K_{n-2} K_{n-1}' \mu_n^{ab}$$

$$\vdots$$

$$+ K_0 K_1' K_2' \dots K_{n-2}' K_{n-1}' \mu_n^{ab^{n-1}}, \quad n \geq 2$$

The extension of those two results to the case where  $x$  and  $u$  are vectors is straightforward.

### References

- BELLMAN, R. *Dynamic Programming*. 1957. Princeton; Princeton University Press
- GUIGNABODET, J. Dynamic programming: Cumulative errors in the evaluation of an optimal policy. *A.S.M.E. Paper* 62-J.A.C.C.-13, June 1963; *J. Base Engng (Trans. Amer. Soc. mech. Engng)*

### DISCUSSION

#### Further Remarks by the Author

When applied to a practical problem, Theorems 1 and 2 rapidly give unreasonably large values of the predicted error bound as  $n$  increases, from which one can draw few conclusions about the true behaviour of the computation errors. In fact, numerical experimentation has shown that a practical envelope of the errors is always a function of  $n$  with a decreasing derivative and thus does not increase nearly as rapidly as  $J_n(\lambda)$  and  $J_n(\lambda, \psi)$ . This is due to the fact that, at some stage of the computation, only a fraction of the largest error at that stage contributes to the largest error at the next stage. This is intuitively obvious and can be given a mathematical interpretation by introducing the coefficients  $\alpha$  and  $\beta$  smaller than one, as shown below. With  $a = b = 1$  and  $X_n = X$  for all  $n$ , Theorem 1 can be rewritten

$$J_n(\lambda) = \alpha J_{n-1}(\lambda) + J_{n-1}(\lambda), \quad n \geq 2$$

$$J_{n-1}(\lambda) = \beta J_{n-2}(K'\lambda) + K K' \lambda, \quad n \geq 3$$

$$J_1(\lambda) = 0, \quad J'_1(\lambda) = K K' \lambda$$

Indeed, we no longer have a true upper bound but a practical way of choosing the quantization interval of the state and control variables, the size of the set  $X_n$  at which  $h_n(x)$  needs to be evaluated, and, eventually, the sampling intervals so as to carry out the computation, keeping the errors within prescribed bounds. The coefficients  $\alpha$  and  $\beta$  can be determined from numerical experimentation for large classes of computational processes.

R. BOUDAREL, *Centre d'Etude et de Recherches en Automatisme, 3 Boulevard Victor, Paris XV<sup>e</sup>, France*

In his very interesting paper devoted to an important problem of error computation, the author considers the case in which the theoretical recurrent equations are replaced by approximations such as

$$h_n^c(x^*) = \max_u F(g(x^*, u), u) + h_{n-1}^c Q(g(x^*, u))$$

where  $Q(g)$  is the quantization operator

Does the evaluation of  $h_{n-1}^c(g(x^*, u))$  by a simple (or a higher order) extrapolation, in spite of frequent discontinuities of functions  $h_n(x)$ , yield more important errors than evaluation by simple round-off?

J. GUIGNABODET, *in reply*

I wish to thank Dr. Boudarel for pointing out the very important problem of interpolation procedures. The 'cost' or 'return' functions  $h_n(x)$  that are encountered in control processes, are usually continuous, and increasing the order of the interpolation surfaces would certainly improve to a great extent the accuracy of the computation. However, such procedure would require a very large number of matrix inversions and cannot be considered in practical cases. The use of orthogonal polynomials to approximate  $h_n(x)$  over the entire range of interesting values of  $x$  seems quite promising whenever polynomial interpolation is itself meaningful. This procedure must be used with great care and the simplest interpolation formulae are often the most satisfactory.

# THEORY OF SELF-ADJUSTING SYSTEMS

## Adaptive Control

A Survey by JOHN G. TRUXAL

In the few years since the founding of the International Federation of Automatic Control, and indeed in merely the three years which have elapsed since the First International Congress in Moscow, the confraternity of control engineers has demonstrated adaptive characteristics far surpassing the most elegant of man-made control systems in speed of response, brevity of settling time, and insensitivity to external noise and changes of personnel. For in that brief period of less than a decade, control engineering has developed from a narrow interest in linear, single-loop electromechanical controllers to a breadth of concern which encompasses the themes of modern stability theory, optimization techniques, adaptive systems, finite automata, learning and pattern recognition, and which touches upon such diverse scientific fields as man's concern with natural physical phenomena on earth or elsewhere, with man himself and the operation of his neural and physiological mechanisms, and with the analysis of man's own social institutions. As a result of this adaptivity of the control engineering species, this Second I.F.A.C. Congress convenes on a non-nationalistic basis with a common desire to exchange information on recent advances in the control engineer's quest for improved technological tools to move ahead in this multiple role in furthering the conquest of space and the understanding of physical, biological, sociological, and economic phenomena.

In this rapid growth of modern control engineering, a central theme has been that of adaptive control: in non-mathematical terms, the control system in which the dynamic characteristics of the controller are purposely designed to vary in accordance with unpredictable variations of the state variables. Here the state variables are to be interpreted as including input signals, environmental variables, and those process variables which indicate the values of process parameters. Furthermore, in the evaluation of the adaptive system, there should be a penalty associated with the failure of the controller to react in accordance with the changes in the state variables.

Such a loose definition of adaptive control clearly encompasses a great many conventional systems as well as systems which are designed and analysed as adaptive. For example, from an appropriate viewpoint, the familiar single-loop system with a saturating forward amplifier is adaptive, since the equivalent gain of the non-linearity depends upon the error signal level (and hence the input signal); furthermore, the quantitative

evaluation of system performance certainly depends upon this non-linearity. In such a case, however, no obvious purpose is served by terming such a system adaptive; we avoid inclusion of such trivial cases by imposing the requirement that the controller be purposely designed to be insensitive to the variations of the state variables: in other words, that the system be designed from an adaptive viewpoint.

While the literature of control theory is replete with arguments about the definition, progress in adaptive control technology has not been impeded by failure of the purists to reach universal agreement on an appropriate definition. On the contrary, the three years since the First I.F.A.C. Congress have witnessed fundamental developments in two quite separate directions:

- (1) The development of a fundamental theory of adaptive control.
- (2) The design of practical adaptive controllers.

As a result of these two distinct directions of research and development, important contributions have been made in four primary areas:

- (a) Control theory.
- (b) Control engineering.
- (c) Control education.
- (d) Control in related fields.

In the following, we outline the nature of the developments in (1) and (2), and then attempt to depict the nature of the contributions (a)–(d).

### Theory of Adaptive Control

In spite of the surge of interest in adaptive control, approximately six years have produced disconcertingly little fundamental theory. Part of the source of the difficulty can be easily demonstrated. *Figure 1* shows the terminology of the fundamental

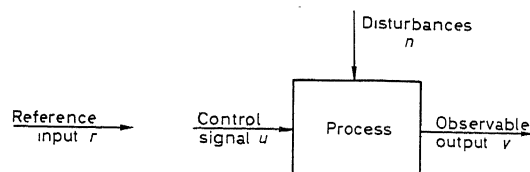


Figure 1. Control system notation

# ADAPTIVE CONTROL

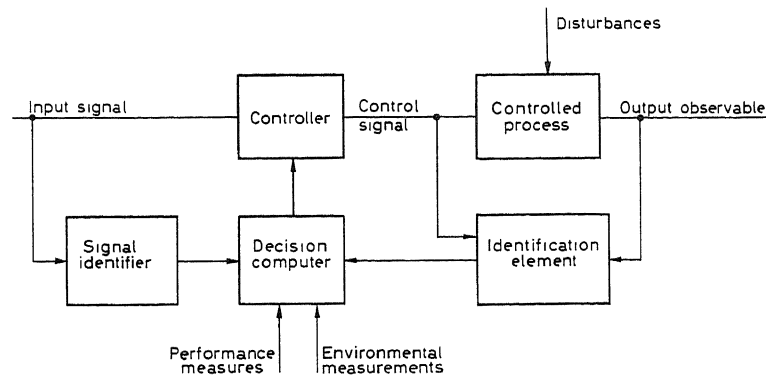


Figure 2. Basic adaptive system

control problem: here  $v$  is the vector representing the observable process outputs,  $u$  the control signal,  $n$  the disturbances, and  $r$  the input signals. The fundamental adaptive system is portrayed in Figure 2: here the process is identified, the input signals are characterized, and the decision computer selects the controller characteristics which yield satisfactory performance during variations of the input, the disturbances, or the process parameters. The problem is normally further complicated by constraints on the allowable values for  $u$  or constraints on the process state variables.

The system of Figure 2 illustrates that a logical theory of adaptive control should be based upon the theories of:

- (a) Identification.
- (b) Optimization.
- (c) Stability.

Unfortunately, the theories in these three areas were notably inadequate six years ago when interest in adaptive control reached an initial peak through the stimulation provided by the problems of aircraft control over a broad flight regime and in an environment which was largely unpredictable and by the growing interest in the analysis of adaptive biological systems with control and communication theories. Consequently, during the last few years major interest in control theory research has focused on the particular areas of identification, optimization, and stability—each an area of major theoretical and practical importance in its own right.

The general optimal control problem was posed by Feldbaum in 1953<sup>1</sup>. Basically the problem is to determine a control input  $u$  which minimizes a given performance index ( $S$ ) subject to certain constraints and given plant dynamics. The plant dynamics, which relate the plant output  $v$  and state variables  $x$  to the control input  $u$ , are usually specified in the form

$$\dot{x} = f(x, u) \quad (1)$$

for continuous time systems or in the form

$$x_{k+1} = f(x_k, u_k) \quad (2)$$

for discrete-time systems.

In 1956 Bellman suggested the application of the theory of dynamic programming to the solution of optimal control problems<sup>2</sup>. This theory leads to a first-order partial differential equation for the solution of the optimal control input. The optimal control input is then obtained as a function of the current state of the system; hence we obtain a feedback structure for the realization of the optimum control system.

At about the same time, a group of Russian scientists under the direction of Pontryagin was developing a theory of optimal control based on the classical calculus of variations—work which has become known as Pontryagin's maximum principle<sup>3</sup>. A solution for the optimal control input in this case involves the maximization, with respect to  $u$ , of the function

$$H = -F(x, u) + p^T f(x, u) \quad (3)$$

where  $F$  represents the integrand of the performance index

$$S = \int_{t_0}^T F(x, u) dt \quad (4)$$

and also involves the solution of a set of  $2n$  ordinary differential equations (where  $n$  represents the order of the state vector  $x$ ). Boundary conditions are of the mixed type, specified both at  $t_0$  and  $T$ .

A third approach to the optimal control problem was pioneered by Kulikowski (1959) in the work on functional analysis<sup>4</sup>. In the United States, the functional analysis approach was developed by Kelley, Bryson, Ho and others under the general title of methods of successive approximations or method of gradients. In this approach, the control input  $u(t)$  is generally obtained as the limit of a sequence of control inputs, defined over the entire optimization interval  $(t_0, T)$ , which result in successively smaller values of the performance index  $S$ . The control input so obtained is generally of an open-loop nature (though a successive approximation technique which results in closed-loop control has been suggested by Bellman under the title of approximations in policy space). Thus, the computational advantages usually associated with this approach must be weighted against the disadvantages associated with open-loop operation, although the advantages of closed-loop operation for the general non-linear system are largely intuitive.

Recently, Kelley, Bryson and Breakwell have suggested gradient techniques to obtain nominal open-loop solutions combined with other techniques (maximum principle, dynamic programming) to obtain closed-loop operation from small deviations from nominal<sup>5</sup>. This approach is very attractive in the design of non-linear systems, since the feedback solution is obtained for a linearized system operating in the vicinity of the nominal solution.

The theory of optimal linear systems<sup>6</sup>, which result when the plant is linear, the performance index is the integral of a quadratic form in  $x$  and  $u$ , and there are no constraints, is well developed both for the stochastic and for the deterministic cases.

In this area of linear control, the concepts of controllability (can the plant be controlled with any input, optimal or not) and observability (can state  $x$  be inferred from measurements of the observable variables  $y$ ) play special roles for multivariable systems<sup>7</sup>. A recent paper by Kalman develops the beginnings of a relation between the optimal approach to linear systems and the classical design methods. It is shown, for example, that in order for a given linear control law to be optimal with respect to some quadratic performance index, the return difference must have a magnitude larger than unity over the entire frequency spectrum. The theory as developed thus far does not, however, permit quantitative consideration of the usual practical problem in which the return difference magnitude cannot be made unity over more than a finite portion of the frequency spectrum.

In addition to this general theory, research in optimum control has considered a variety of special problems—e.g., the minimum time problem<sup>8</sup> (the evaluation of that control signal which effects a desired change in system state in the minimum time), or the vehicle control problem in which minimum final error is desired when the problem is complicated by uncertainties in observation of vehicle position and velocity and by limited total available quantities of fuel<sup>9</sup>.

In the light of this general optimization theory, therefore, the adaptive control problem relates to the design of the optimum system in the presence of unpredictable variations in the state variables and with the controller designed to track, in some meaningful way, these variations. In the most general approach, the adaptive and optimizing aspects of the problem are not separated, but rather the problem is considered as a single entity: the optimum control of a stochastic or unknown process. In the linear case in which the state variables are corrupted by noise, for example, the solution is known to be an optimum estimation of the state variables from the observable variables, followed by the controller designed on the assumption that the state variables are known precisely. In the more general case, the control signal  $u$  must perform the dual function of effecting the desired control of  $y$  and simultaneously permitting continually improved estimation of the process parameters.

The first extensive study of discrete-time stochastic control was published by Feldbaum in 1960; studies of continuous-time stochastic control problems appeared in the following year<sup>10</sup>. In all these cases, the complexity of the solutions realized seems, at least at the present time, to restrict applicability severely. It is generally assumed by the above authors that the stochastic variables introduced in the control system result in a Markov process for the state variable  $x$ . When dynamic programming is applied to the continuous-time stochastic control problem, a second-order partial differential equation results.

In an attempt to simplify the theory and bridge the gap between general mathematical solutions and practical, engineering realizations, the optimization and identification problems have been considered separately. Identification is of importance in its own right as a fundamental element of control system design, since feedback control is primarily important because the process parameters are unknown precisely or vary with time or environmental parameters. Identification for adaptive control differs from the classical problem of network and control theories, however, because of the emphasis on minimizing the time required to complete the determination of significant process characteristics.

Most intensive studies of the identification problem have related to the evaluation of the impulse response of the process, or the matrix of impulse response functions (just as most practical attempts to design adaptive controllers have operated from estimates of the changes in the various impulse response functions). Work has focused on the minimum time to determine the impulse response with smoothing of the measured data to procedures for the determination of the maximum likelihood estimate—in both cases, with particular emphasis on the problem depicted in Figure 3<sup>11</sup>.

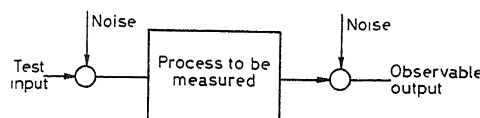


Figure 3. Identification problem

Attempts to convert adaptive control theory to practical realizations have generally utilized a partial identification: e.g., determination of the damping ratio and/or the resonant frequency for a system known to be dominantly second-order, measurement of only the early portion of the impulse response corresponding to the period of time over which control is to be realized by the last control signal (with subsequent control signals designed to override the effects of earlier signals and the later portions of the impulse response), or determination of only such gross features as the variation of the approximate duration of the impulse response or the settling time of the process. In such cases, which represent a segment of the bridge between the general theory and practical engineering realizations, very little theoretical analysis of accuracy, speed of identification, and other aspects seems to have been reported in the control literature.

Just as the identification problem can be severed from the theory of adaptive control for analytical and design purposes, the stability considerations can be separated from the design of the adaptive loop. Until the interest in adaptive control arose six years ago, the control engineer was predominantly concerned with the classical stability analyses of linear, time-invariant systems according to the familiar graphical and analytical criteria of Routh, Hurwitz, Nyquist, *et al.*, or with the analysis of non-linear systems via the phase portrait, the analysis with describing functions or other approaches to quasi-linearization, or direct linearization procedures.

At approximately the same time as interest arose in adaptive control, Liapunov stability theory came to the attention of the control engineer, and serious efforts were made both to analyse and to design non-linear control systems of practical engineering importance<sup>12</sup>. While a variety of the optimization procedures which have been developed lead inexorably to stable systems through the choice of a performance measure which results in satisfaction of the Liapunov conditions, the adaptive systems which have been built often are described by such complex non-linear differential equations as to discourage any attempts at a theoretical study of stability. Extensive analogue and/or digital simulation and testing provide the only available, albeit mild, assurances of system stability. The difficulties arise, of course, because of the time delays associated with identification (whether or not the identification portion of the system be separated from the controller), the multiple non-linearities represented by the constraints and by the adaptivity of the controller, and the time-

varying nature of the system—all of these characteristics combined with the high order of most realistic systems of engineering interest. While studies of simplified approximations to specific, actual systems have indicated in several cases the possibility of instability, the prospect is certainly not hopeful for meaningful studies of actual systems; it appears that stability theory, as it develops during the next three years, will primarily be useful for the increased understanding of the dynamic properties of relatively simple adaptive systems.

Thus, the theory of adaptive control is evolving, built largely upon the three distinct areas of optimization, identification, and stability. A fourth focal aspect, that of sensitivity, although of apparently equal significance, has been developed only superficially. Several facets of the sensitivity problem are illustrated by the simple adaptive system of *Figure 4*. Here the controller is varied automatically to minimize a performance measure in the presence of slow variations of the process parameter  $g$ . (The  $g$  may be a particular physical parameter or a derived characteristic, such as the deviation of the process from a reference model or a relative damping ratio.)

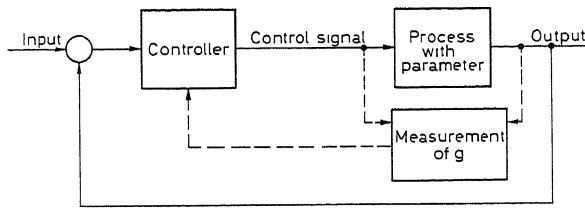


Figure 4. Simple adaptive system

Two quite different aspects of the sensitivity are of interest<sup>13</sup>. First, we may consider the sensitivity of the overall system transmission (or the performance measure) to the changes in  $g$ ; in such a case, the adaptivity can be interpreted as a technique for reduction of this sensitivity, and design and evaluation of the adaptive loop can be effected from this sensitivity viewpoint. Alternatively, we may be interested in the sensitivity of system performance to changes in other process or environmental parameters—as a measure of the overall merit of the adaptive system. In both cases, the close relations between sensitivity and stability for linear systems (the relations which indicate the parameter margins for the system stability, or how much specific parameters can vary before the system breaks into oscillation) provide a basis for stability studies in the presence of parameter variations.

### Adaptive Controllers

Within the realm of practical adaptive control, we include those design approaches which focus on the concept of improved system performance, and which attempt to realize this goal by restricting consideration to a sub-optimal problem of system design. For example, within this category fall those systems which operate on the basis of the adjustment of a single parameter (or a small number of parameters) to extremalize a selected function of system performance. The system configuration and the individual elements are selected according to conventional control techniques, after which the optimization is included in order to improve performance within the framework of the original system design.

The technical literature of feedback control is replete with examples of such sub-optimal systems developed during the last few years. In general, such designs are characterized by simplicity both conceptually and practically, reliability, and emphasis on practicality. In terms of control systems theory, such designs represent major contributions primarily in terms of the novel configurations which results—systems which never would evolve from the conventional control theory focusing so heavily on stability considerations for linear, single-loop feedback configurations.

The research efforts along this direction of specific adaptive systems can be illustrated by three specific approaches, the first of which is depicted in *Figure 5*, a sketch of the M.I.T. Instrumentation Laboratory adaptive system which has evolved from the early work by Draper and Li on optimalization. In this configuration, a conventional feedback control system (represented in simplified form by the single control loop in the figure, even though in most cases the configuration is multi-loop) is designed; superimposed on this is the optimizing sub-system. The actual system output is compared with the response of a model to yield a generalized error  $e_g$ . A non-linear function of this  $e_g$  is used to adjust a parameter  $K$  of the control loop to yield a minimum of the performance measure.

Certain basic problems arise in the design of such a model-reference adaptive system:

- (1) How is the model to be chosen?
- (2) How are we to select that parameter  $K$  which is to be varied?
- (3) How is the performance measure or criterion to be chosen?
- (4) Under what conditions is the adaptive loop itself stable?

While certain aspects of these problems have been investigated<sup>14</sup>, the extension of this approach to complex configurations relies

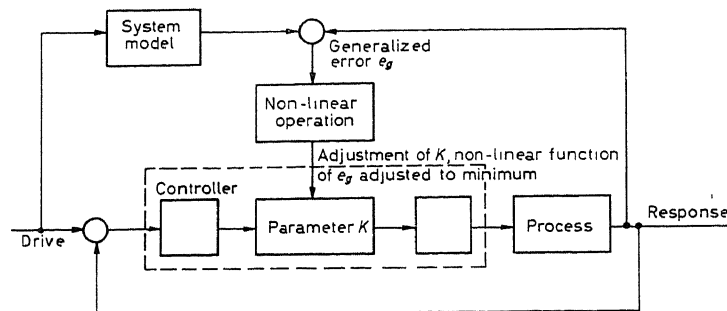


Figure 5. Model-reference adaptive system

heavily on the final design and verification of performance characteristics via simulator studies. Such a situation should be neither surprising nor discouraging, however, in view of the fact that optimization techniques are particularly useful in the design of complex systems in which the engineer is faced with the difficult question of how to start.

The extensive development of the configuration of Figure 5 has been motivated, at least in part, by the adaptive autopilot problem for piloted aircraft moving through radically varying environments—a specific engineering problem which also stimulated the development of the Minneapolis-Honeywell autopilot configuration<sup>15</sup>. In this system, the controller gain is adjusted automatically to maintain a small-amplitude limit cycle of a suitably chosen process variable. In more recent work, a digital computer is utilized as the controller to provide digital compensation, to furnish the required forward loop control, and to simulate the automatic gain changer, the model, and the gyro blender and circuitry which is used to control the body-bending modes. This particular system typifies the strong trend today toward digital realizations of adaptive controllers.

Complementary to the aircraft industry's interest in adaptivity has been the emphasis on industrial applications of computer control: once the computer is installed for data logging and routine processing of system performance information, for shutdown, start-up and emergency control, and for the automation of routine economic and profit analyses, the systems engineer visualizes utilization of the computer flexibility to realize optimum plant performance, or at least control which is more precise and faster than that achievable with human operators.

In the technical literature describing applications of optimum computer control, two approaches dominate: the model approach and the automatic experimental approach<sup>16</sup>. Under the former design philosophy, a model of the process is used to determine the optimum control signal as a function of measurable signals and disturbances. The response can be very rapid, can avoid difficulties with multiple extrema, and can include learning or updating modification in the model; on the other hand, performance is limited by the accuracy of the model, and realization of performance near optimum requires intensive studies of and measurements on the process to be controlled.

In the automatic experimental approach, the optimum is realized by the injection of artificial input signals to perturb the operating point, with a subsequent evaluation of whether the performance improved or deteriorated as a result of the change. In such an approach, the speed of response is limited by the fact that the system must search for an optimum, the presence of multiple extrema causes difficulties, and the existence of many signal variables leads to exceedingly slow and possibly poor performance. A considerable portion of the optimization literature of the past few years is devoted to study of searching techniques to overcome one or more of these difficulties.

Chen and Decker emphasize the advantages to be gained by combination of these two approaches, with the performance of the composite system indicated in Figure 6. The plot shows the payoff  $J$  as a function of the control signals  $u$  for various disturbance inputs ( $d_1$  and  $d_2$  here). The solid curve portrays the variation of  $J$  with  $u$  for the actual plant, the broken curve the corresponding solution for the simplified model of the plant. The two constraints indicate the allowable bounds on  $u$  and  $J$

which result from considerations such as safety or which are imposed to avoid subsidiary extrema.

If the system is initially operating at point 0 with a disturbance input  $d_1$ , and the dynamic performance is demonstrated by a change from  $d_1$  to  $d_2$ , we find that the performance moves initially to point  $E$  since  $u$  cannot change instantaneously. As fast as the model system responds, operation moves to point  $C$ ; thereafter the automatic experimental procedure moves the system toward the optimum operating point  $B$ . Clearly we are using the model here for fast, gross corrections, the automatic experimental system for the slower, fine corrections, although the specific interrelation of these two portions of the system may be quite complicated in problems more realistic than the simple situation depicted in Figure 6.

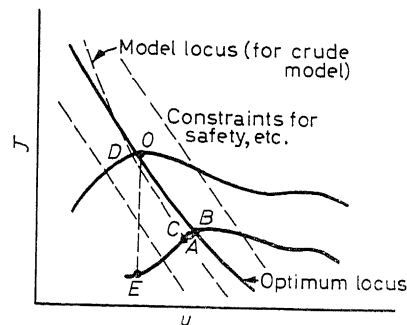


Figure 6. Example of restricted optimization and adaptivity

The final aspect of restricted and simple adaptive control to be discussed is based upon the utilization of simple digital logic for the determination of the desired control signal  $u$  on the basis of inspection of the response of a high-speed model of the process. If for conceptual simplicity we assume the discrete case, we can consider the problem of selecting the control signal which is to be constant over each interval of time. At the time  $t = 0$ , we wish to determine the  $u$  for the first interval from 0 to  $T$ . This determination must be derived from the available information: the present and past values of the control signal, the response  $x$ , and the reference input  $r$ .

The determination can be implemented in the following way. We consider only the values of  $r$ ,  $x$ , and  $u$  every  $T$  seconds (i.e., at the sampling times). On the basis of the known statistical characteristics of  $r$ , we can estimate  $r_1, r_2, \dots$ . On the basis of the past values of  $u$  and  $x$  and our knowledge of the process dynamics, we can determine the response values into the future ( $x_1, x_2, x_3, \dots$ ) for any assumed sequence of control signal values  $u_0, u_1, \dots$ . If we wish to minimize the mean square error, we might attempt to select  $u_0$  in such a way that (assuming  $u_1, u_2, \dots$  are later selected optimally) we would minimize the summation

$$(r_1 - x_1)^2 + (r_2 - x_2)^2 + \dots$$

It seems apparent that the choice of  $u_0$  has a decreasing effect on the successive terms in this series; furthermore, because of the increasing difficulty of predicting  $r_j$  as we look farther into the future, we should weight more heavily the early terms. Such considerations suggest the consideration of only a small number of these terms, for example three:

$$J = (r_1 - x_1)^2 + (r_2 - x_2)^2 + (r_3 - x_3)^2 \quad (5)$$

Here  $r_j$  are predicted values of  $r$ ; the  $x_j$  are future system response values which incorporate the effects of past signals plus the

effects of the variable or controllable future signals  $u_0$ ,  $u_1$ , and  $u_2$ . The optimization involves determination of the minimum of  $J$  over the allowable range of values of  $u_j$ , but we actually apply to the process only the signal  $u_0$ . When  $t = T$  (and  $u_1$  is to be applied), the new optimum value for the control signal is to be re-evaluated.

The attractiveness of this approach to sub-optimization derives from the possibility of simplification in the evaluation of  $u_0$ . If we consider a binary control signal (with  $u$  either  $-1$  or  $+1$ ), we must choose between these two values on the basis of minimization of  $J$ , where we assume  $u_1$  and  $u_2$  will subsequently be selected in an optimum manner. In terms of the logic required to implement this decision on  $u_0$ , we need to divide our three-dimensional space into two parts: one requiring a  $u_0$  of  $-1$ , the other  $+1$ . Our present location in this three-dimensional space is determined from the three predicted values of the future system input and the response with all possible control signals.

Major further simplification is possible if the system is linear, so that the future response can be divided into two, additive parts resulting from past inputs and from future inputs<sup>17</sup>. Then our present location in this three-dimensional space is determined by the predicted values of the future system error with no control input. Implementation of this control scheme involves only high-speed prediction of  $e_1$ ,  $e_2$ , and  $e_3$  without control, and then (by simple digital logic) determination of the location of this state with respect to the division of state space into the two parts corresponding to the two possible control signals. Thus, this system involves only a high-speed model of the process to predict future response values from the present energy storage conditions, a predictor to act on the present and past input, and simple digital logic to determine in which half of the three-dimensional space we are situated.

The three schemes described in this section are only three of a wide variety of practical system realizations derived with emphasis on a sub-optimization approach to adaptive control. In each case, design of the optimizing equipment requires at least a reasonable estimate of process dynamics; in each case, the optimizing system is designed to correct for slow variations or the effects of low-frequency disturbance inputs; in each case, the optimizing components are sufficiently simple to permit simultaneous realization of control over reliability. These advantages are at least to some extent offset by the fact that meaningful analysis of the three systems is apparently not possible if we are working with anything other than the most elementary processes; actual design and verification of the value of the optimization and adaptivity must rest upon computer simulation studies and actual equipment tests.

### Contributions of Adaptive Control Research

In these final paragraphs, it is appropriate that we summarize briefly at least some of the contributions of the intensive work in adaptive control theory during the three years since the First I.F.A.C. Congress. Obviously, it is not possible in such a discussion to separate the work in adaptive control from that in stability theory, optimization, identification, sensitivity, and so forth, since these various aspects of control theory are so totally interdependent.

First, what are the fundamental contributions to control theory? Optimization theory provides a fundamental scientific

basis for control system design. The optimum design serves as a yardstick for quantitative evaluations and interpretations of all of the conventional control system design techniques. In a wide variety of linear problems, optimization theory provides a straightforward approach to system design. In an entire class of problems, for example, the theory indicates that the optimum controller is a saturating amplifier and specifies both gain and saturation level. More generally, the easily determined saturating amplifier may require an associated, linear, dynamic controller; even when these linear transfer characteristics are easily determined only very approximately, the theory indicates the very important structure of the optimum solution. In other problems, the selection of a suitable Liapunov function as the integrand of the performance measure leads directly to a design criterion and solution, with the latter inevitably stable.

Thus, optimization research and the consequent adaptive research have provided an entirely new group of theoretical solutions. In addition, and perhaps of equal importance, the research has focused attention on an entirely new set of concepts. Three years ago we were only vaguely characterizing systems in terms of controllability and observability, a vagueness which carried over from intuitive understanding based upon earlier network and feedback theories. Interest in adaptive control is leading to extensions of sensitivity theory and has generated an interest in identification which transcends the narrow and highly restricted concern earlier of the communication engineer and the control engineer.

Perhaps the most important contribution of adaptive control has been the wealth of entirely new systems which have evolved from the research—systems based on digital logic or the entire group of model-reference configurations: control systems which can never be derived from the 'classical' control theory of 1950. The occasional arguments we still hear today from the 'classicalists' who claim that adaptive control can achieve nothing unobtainable with conventional feedback theory lose sight of the essential nature of engineering: the continual search for novel design approaches. The availability of three solutions to any engineering design problem is always better than having only two; adaptive control has opened up a host of new, unconventional, and exciting solutions.

Any overall evaluation of the significance of adaptive control research must recognize the impact of this research on engineering education and the resulting changes, both today and in the future, in the capabilities and breadth of interest of engineers. In particular, control research and engineering emphases have been primary factors in the early introduction into educational programmes of the concepts of state models and numerical methods. In the former category we should include the relations among state models, transfer functions, flow-graph representations, and analogue and digital simulations—relations which unify so much of system analysis, including (as an example) the various approaches to linear-system stability. In the latter category, we must encompass digital simulation and computer control, as well as numerical techniques for the solution of typically complex analysis, evaluation, and identification problems.

Finally, the interest in adaptive control reflects most poignantly the interrelationships between control engineering and such areas as economic system analysis, pattern recognition, learning theory, medical engineering, and bioelectronics. The term 'adaptive control' was first borrowed from the biological



field; the control engineer's interest in the engineering and scientific theories of adaptive control is paralleled by the biologist's and physiologist's growing interest in the control field.

The author is deeply indebted to the other members of the control research group at the Polytechnic Institute of Brooklyn, and particularly to Professors R. Drenick, P. Dorato, and L. Braun, Jr. The support of the U.S. Air Force Office of Scientific Research under Grant AF-AFOSR-62-280 is gratefully acknowledged.

## References

- The following list is not intended as an evaluated or selected bibliography of work in adaptive control, but rather as a representative group of references, in many cases with extensive bibliographies.
- <sup>1</sup> FELDBAUM, A. A. Optimal processes in automatic control systems. *Automat. Telemekh.* 14 (1953); an approach to the problem of the synthesis of optimal automatic control systems. *Proc. II All-Union Conference on the Theory of Automatic Control*, vol. 2 (1955) *Izd. Akad. Nauk. S.S.S.R.*
  - <sup>2</sup> BELLMAN, R. On the application of the theory of dynamic programming to the study of control processes. *Proc. Symp. Non-linear Circuit Analysis*. Polytechnic Inst. of Brooklyn, Brooklyn, N.Y., April, 1956; also *Adaptive Control Processes: A Guided Tour*. 1961. Princeton, N.J.; Princeton Univ. Press
  - <sup>3</sup> BOLTJANSKII, V. G., GAMKRELIDZE, R. V. and PONTRYAGIN, L. S. Towards a theory of optimal processes. *Dokl. Akad. Nauk.* 110 (1956) 1
  - ROZONER, L. E. Pontryagin's maximum principle in the theory of optimum systems, I, II, III. *Automation and Remote Control*. Vol. 20. 1960 (original Russian in Oct.-Dec., 1959) 1288, 1405, 1517
  - <sup>4</sup> KULIKOWSKI, R. On the synthesis of adaptive systems. *Bull. Acad. Pol. Sci.* (1959)
  - KELLEY, H. J. Gradient theory of optimal flight paths. *J. Amer. Rocket Soc.* 30 (1960) 947
  - HO, Y. C. A successive approximation technique for optimal control systems subject to input saturation. *J. Basic Engng.* 84 (1962) 33
  - <sup>5</sup> KELLEY, H. J. Guidance theory and extremal fields. *Trans. Inst. Radio Engrs*, AC-7 (1962) 75
  - BREAKWELL, J. V. and BRYSON, A. E. Neighboring optimum terminal control for multivariable nonlinear systems. *SIAM J. Control*, 1 (1963)
  - <sup>6</sup> KALMAN, R. E. and KOEPCKE, R. Optimal synthesis of linear sampling control systems using generalized performance indexes. *Trans. Amer. Soc. mech. Engrs* 80 (1958) 1820
  - KALMAN, R. E. When is a linear control system optimal? *Tech. Report* 63-5, *RIAS*. March 1963
  - <sup>7</sup> KALMAN, R. E., HO, Y. C. and NARENDRA, K. S. Controllability of linear dynamic systems. *Contributions to Differential Equations*. Vol. 1. 1962. New York; Macmillan
  - GILBERT, E. G. Controllability and observability in multivariable control systems. *SIAM J. Control*, 1 (1963)
  - <sup>8</sup> SARACHIK, P. E. and KRANC, G. M. Optimal control of systems with multi-norm constraints. *Automatic and Remote Control*. 1963. London; Butterworths. Munich; Oldenbourg
  - NEUSTADT, L. W. Synthesizing time optimal control systems. *J. Math. Anal. and Appl.* 1 (1960) 484
  - <sup>9</sup> ATHANS, M. Minimum fuel control of second-order systems with real poles. *Proc. 1963 JACC, AICHE*, June 1963, Paper IX-3. Minneapolis, Minn.
  - <sup>10</sup> FELDBAUM, A. A. Dual control theory I, II, III, IV. *Automation and Remote Control*, 21-9, 1240-1249 (1960); 21-11, 1453-1464; 22-1, 3-16; 22-2, 129-143 (1961)
  - FLORENTIN, J. J. Optimal control of continuous-time Markov processes. *J. Elec. Con.* 10 (1961) 473
  - <sup>11</sup> LEVIN, M. J. Estimation of a system pulse transfer function in the presence of noise. *Proc. 1963 JACC, AICHE*, Paper XVI-3, June 1963. Minneapolis, Minn.
  - <sup>12</sup> KALMAN, R. E. and BERTRAM, J. E. *Trans. Amer. Soc. mech. Engrs*, Series D (1960) 371
  - <sup>13</sup> The I.F.A.C. Technical Committee on Theory has sponsored a symposium on analysis sensitivity analysers with the Yugoslav National Committee of I.F.A.C. in Dubrovnik on August 31 to September 1, 1964
  - <sup>14</sup> LI, Y. T. and WHITAKER, P. Some research work in self-adaptive systems in M.I.T. aeronautical and astronautical eng. dept. *Proc. I.F.A.C. Symposium on Optimizing and Adaptive Control*. June, 1963. Pittsburgh, Pa.; I.S.A.
  - <sup>15</sup> LEE, J. F. L. A digital adaptive flight control system for flexible missiles. *Trans. 7th Symp. Ballistic Missile and Space Technology*. Vol. II, Aug. 13-16, 1962, 115-148. Los Angeles, Calif.; Air Force Systems Command and Aerospace Corp.
  - <sup>16</sup> CHEN, K. and DECKER, R. O. Process optimization by combining the model and experimental approaches. *I.S.A. Trans.* 3, July 1962. 279-285
  - <sup>17</sup> HORING, S. On the optimum design of predictor control systems. *Automatic and Remote Control*. 1963. London; Butterworths. Munich; Oldenbourg



# Learning Machines

A Survey by GORDON PASK

## Acknowledgements

*I would like to thank Mr. F. Coales, Dr. A. M. Uttley and Professor J. H. Westcott for their helpful criticism, and Mr. J. D. Cowan and Mr. B. N. Lewis for reading and commenting on the manuscript. My own work in this field has been mainly supported by the Air Force Office of Scientific Research, under Contract AF61 (052)-640 through the European Office of Aerospace Research (OAR), United States Air Force.*

## Summary

Learning machines have increasing practical importance and it is necessary to develop a competent theory for these systems rather than treating them as though they were simply an intractable kind of adaptive controller. Since learning is a goal directed change in a pattern of behaviour, it is a property that depends upon the environment and the objective of the machine, as well as the specification of the machine itself. Briefly, it is a property of an organization and, consequently, cybernetics offers the most suitable framework for a theory of learning machines.

Present-day learning machines, which may be large and elaborate networks of adaptive components or the equivalent computer programmes, stem from the pioneer work upon conditional probability machines and homeostatic systems. The mechanical concomitant of learning always involves structural modification, over and above adaptive change in the system parameters, but the modifying process, although it may involve 'random' variation, is well specified in any realistic device (purely 'random' search is impracticable). Further, a useful learning machine has an hierarchical organization. If it is part of a control system there will be different levels of goal associated with different levels of control. In any case there will be more or less abstract representations of the environment. When the machine is required to learn about a symbolic or linguistic environment (in problem solving or in man-machine symbiosis) these levels correspond to metalanguages in terms of which internal data processing is carried out and external communication can readily take place. An evolutionary learning machine is able to build its own hierarchical structure and in some conditions this process may be identified with a form of 'concept' acquisition. In order to describe the underlying mechanism, it is necessary to introduce the idea of 'reproducing' or unlocalized automata.

Where design principles are concerned, brains and neural structures still provide the most useful data, though there is no necessary relation between the structural or functional specification of a brain and of a learning machine. Certainly there are features of development, like the need for maturation in contact with the environment that are common to each system. On the assumption that the biological analogy contains a fair grain of truth, the realization of tangible learning machines will call for large flexible and computationally 'parallel' assemblies of coupled transmission lines. It has been proposed that these active transmission lines would be optimally embodied in macromolecular structures and this suggestion appears to be physically plausible.

## 1. Introductory Comments

### 1.1. Importance of Learning Machines

Learning machines are more than the ingenious toys they are commonly supposed to be. They constitute a large and important

class of automata. Members of this class that have been realized as artefacts (mechanical or electronic models) provide existence proofs, often the only existence proofs available at the moment, for propositions in the theory of learning machines. The theory of automatic control and finite automata (devices such as switching networks that compute a function of a well defined input) is a beautifully constructed but highly specialized fragment of the theory of learning machines.

Beyond a certain limit, the fragment of a theory cannot be usefully extrapolated. Conceptually plausible analogies with familiar constructions break down and it becomes more convenient to tackle the problems that arise in terms of a framework that embodies the original and well-known area of study as a special part. My contention is that this critical stage has been reached in connection with many problems of automatic control. Attempts to describe essentially learning situations in terms of the conventional formulations are often strained, or inadequate, or beset with illegitimate approximations. Whether he likes it or not, the control engineer is being forced to accept the need for a theory of learning machines and not infrequently he is also required to develop a theory where none exists.

In this case, learning machines, however crude they may be, provide valuable indicators (some of them suggest solutions to particular problems; for example, in pattern recognition), but the mere existence of learning machines is insufficient. The present theoretical background (which is firm but far from elegant) needs rapid development in order to serve a new cybernetic technology of control and computation in systems previously regarded as intractable.

To restrict the discussion, obviously trivial learning machines (tape recorders with cunningly manipulated inputs and outputs) will be excluded even though they are presented in a plausible guise. Next, non-trivial learning machines will be excluded if they are more opportunely described in some other fashion (adaptive systems like Gabor's<sup>1</sup> genuinely 'learn' about a restricted environment but their activity can be analysed within the conventional theory of control and 'learning' need not be considered). One can omit the field of finite automata, in which the work of McCulloch and Pitts<sup>2</sup> on neural network analogues is pertinent to brains, but oriented towards computation rather than learning. Finally, for lack of space rather than relevance, a detailed examination of the 'learning algorithms' that feature in 'artificial intelligence' is avoided although it will be necessary to view the field rather broadly in 2.8. There are still, to my knowledge alone, over 350 significantly different learning machines which have either been simulated by computer programmes or constructed as special mechanisms.

### 1.2. Deterministic and Probabilistic Systems

A system, whether it is an animal or a machine, is always specified in relation to a particular method of observation. In the case of animals this specification usually amounts to the experimental conditions, the environment in which the animal is

placed (a rat in a maze) and the measurements it is proposed to make (whether the rat turns left or right). In the case of a machine the specification is often dictated by the circuitry and, in so far as control mechanisms are concerned, the environment to be controlled. At any rate, whenever a system is mentioned there is also, by definition, a state description that lists the relevant states, which, in the ultimate analysis, are states which, for often excellent reasons, we, the experimenters, deem relevant.

A behaviour is a sequence of output states (or actions) that are contingent upon input states or evidence and upon the internal state of the system. If a behaviour is repeatable and consistent it can be ascribed to a particular computation that the system performs (in other words, in view of the observed regularity, the system can be described as though it had been built to compute a given function of its input). Of course, if the system is a machine that has been built for some purpose, it will be possible to infer this from an inspection of its circuit. Statistical invariants are countenanced as regularities and if these exist, even in the absence of any deterministic pattern of behaviour, the system is said to compute a probabilistic function of its input.

There is partial ignorance of the internal state and the exhaustive input specification of men and animals, their experimental gambits are only defined at the probabilistic level. For constructed machines, with the exception of some evolutionary systems introduced in 2.7, a deterministic specification is possible. It may, however, be completely impracticable, because the system is very large (networks in 2.4), because there is a 'redundancy of computation' so that the same easily recognized behaviour may be due to the action of different mechanisms (as in many stable 'error correcting' computers), or because internal parameters of the system are changed as a function of its previous states in such a way that it proves impossible to assume a momentarily static structure.

Adaptive machines with a 'black box' for providing an uncorrelated forcing input that modifies their parameter values, for example, the adaptive controller in *Figures 6 or 7* and some of the machines described in 2.3 and 2.4 are strictly deterministic if the interior of the 'black box' is examined but it is useful to describe them as though they were probabilistic devices if the sequence of values of the forcing input is irrelevant (as it is when the machine is intended to control a statistically defined environment). Nor are 'random' learning networks (like those described in 2.4) really probabilistic, because their initial connectivity is determined by a particular 'random' number table. In this case, however, a very different probabilistic fiction is sometimes useful. The 'random' number table is chosen to have certain statistical constraints and the particular network is regarded as a 'random' sample from a statistical ensemble characterized by these constraints. Whereas the sequence of input events was previously deemed irrelevant to the adaptation of the given machine, in this case the variations of the machine within the statistical ensemble are deemed irrelevant.

### 1.3. Minimal Criterion for Learning

A minimal criterion for learning is a goal directed change in computation or possibly 'a novel computation'. By analogy, when one says that a man 'X' has learned a skill or learned a telephone number one does not merely mean that some change has occurred in his brain (it certainly has, of course, and some kind of adaptation is a prerequisite for learning). One also means

that the man has acquired the computing process involved in performing the skill or reciting the telephone number.

Further, since learning is a property of a system's behaviour, an immediate inference from 2.2 is that it will depend upon the experimental conditions (this machine will be a learning machine in some circumstances but not in others). The man 'X', for example, can only be said to have learned if he is presented with an environment in which he can recall his skill or his telephone number.

Finally the property of learning is unaltered in an isomorphic model or representation of the system since it does not directly depend upon material construction. Nobody argues that men learn because they are made of protein or that computers cannot learn because they are made from metal. Rather, 'learning' belongs to an organization some part of which may be common to men and computing machines. To summarize these points one can say that 'learning' is a cybernetic construct and is capable of embodiment in cybernetic models that represent organizations. The condition that learning is goal directed implies that an observer or model builder can understand the change in computation that takes place as a result of learning and restricts the discussion to non-trivial cybernetic models.

In fact, the most interesting kinds of learning involve more specialized changes than have been described. A system may learn to abstract features of its environment (in other words, to represent a class of events in terms of attributes of this class). It may associate equivalence classes of its internal states (called 'signs') with subsets of points in a descriptive space which has these features or attributes as its coordinates, when the 'sign' may be called a symbol and the subset of descriptive points, its denotation. The system may combine symbols or form classes of invariant relations between symbols (which, in the logician's sense, can serve as universals). Finally, one might (in some carefully selected circumstances) refer to the process whereby internal or external events are signified and manipulated as 'percepts' and 'concepts'.

A great deal of machine learning does involve processes that appear to be isomorphic with 'concept acquisition' (using this phrase like those objectively oriented psychologists who aim for a functional account of mentation). Whether a given observer is prepared to say that the machine concerned does acquire concepts, or, indeed, whether he is prepared to say it learns (rather than 'learns') is a matter, as MacKay<sup>3</sup> points out, for that observer to decide. The decision has very little to do with the manifest behaviour and chiefly depends upon his personal view about how much he resembles the machine<sup>4</sup>. All a cybernetician can assert is that for some systems there is no mechanical absurdity involved in the contention that the device is able to learn or conceptualize. So far as the minimal criterion is concerned, at any rate, there is a tacit assumption that a 'learning' machine is literally accepted as a learning machine.

### 1.4. Machines that Imitate Organisms

There are two broad categories of learning machine:

- (1) Devices that imitate the undeniable learning behaviour of an organism.
- (2) Devices that manifest the abstract cybernetic property of learning. Frequently, the apparatus must also satisfy a practical goal like learning to recognize distorted alphabetical characters or learning to control a non-stationary process.

In either case, a simulation (a tangibly realized cybernetic model) must embody salient features of the environment as well as the learning system itself.

Members of category (2) are fairly well known and are considered at length in 2.2. For the moment let us examine a few members of category (1), which are less familiar in the discipline of engineering.

The learning machines of (1) simulate animals or brains or parts of brains. Typical small animal simulators are Deutsche's rat<sup>5</sup> (which 'learns' to run a maze according to the tenets of Deutsche's learning theory) and Angyan's turtle<sup>6</sup> (which 'learns' about a more liberally specified environment, using the strategies of sequential conditioning). In each case, the behaviour of a real animal is abstracted and the inbuilt mechanism constitutes a psychological or neurophysiological hypothesis about the generation of learning behaviour which is tested and explicated by embodiment in an artefact. From the viewpoint of control engineering, their chief value stems from the argument given in 3.1, namely that although there is no necessary connection between learning machines and brains or animals, the most potent clues about the design of learning machines come from biology.

The first of these models was constructed by Grey Walter<sup>7</sup> in the form of a mechanical 'tortoise' with auditory, visual and tactile sensory inputs and motor capabilities that drove it around the floor with a behaviour initially determined by inbuilt light responsive reflexes. Grey Walter used this machine to demonstrate the acquisition of conditioned reflexes. In particular, he aimed to show the consequences of a four stage learning hypothesis.

(a) Signals representing unconditioned stimuli (in other words, stimuli that already give rise to some response, due to built-in association) are abrupt and demark the onset of a stimulus, whereas signals representing novel stimuli (which may become conditioned stimuli) are 'stretched'.

(b) The overlap between unconditioned and (already 'stretched') novel stimuli is counted as a measure of the regularity of this joint event.

(c) If the overlap count exceeds some threshold value an autonomous process, such as a damped long term oscillation, is initiated. Very tentatively, this event corresponds to an 'insight'.

(d) If the novel stimulus is combined with the existence of the autonomous oscillatory process that indicates frequent coincidence between the unconditioned and the novel stimuli, the novel stimulus becomes coupled to and will elicit the existing unconditioned response. Hence, the novel stimulus becomes a conditioned stimulus.

Learning machines that simulate the activity of a brain differ according to the degree of physiological verisimilitude and the level of abstraction involved in their design. At one extreme there are models, like Harmon's artificial neurones<sup>8</sup>, that accurately replicate the characteristics of a real neurone (in so far as these are known from microelectrode recordings). Small collections of these elements (corresponding, at most, to ganglionic organizations) exhibit versatile behaviour patterns, some of which have properties akin to learning. At the other extreme there are statistical models, such as that of Beurle<sup>9, 10</sup>, in which a large collection of artificial neurones, characterizing a more abstract image of the real neurone than those designed by Harmon, are connected to form a network that replicates the statisti-

cally defined connectivity of the mammalian cerebral cortex. If one of Beurle's self-oscillatory models is coupled to a suitable environment it, also, manifests learning behaviour.

At this level, the systems of (1) and (2) above are distinguished by intention alone and most of them are open to a dual interpretation. Thus one of the conditional probability machines of Uttley<sup>11-15</sup>, which are described in detail in 2.3, can be arranged to imitate a brain very closely, but its value does not depend upon imitation. A conditional probability machine is interesting in its own right.

### 1.5. *The Material and Symbolic Dependence of a Realizable and Effective Learning Machine*

Although the fabric of a learning machine, synthesized on the basis of a cybernetic model, is irrelevant to learning (so that there is no need to distinguish between a computer programme and a special purpose device) it is not true that naturally occurring machines (animals and other organisms) are independent of their material character. The natural and the synthetic machines are both said to learn because their computation changes, but an animal changes its behaviour in order to survive in a changeful environment. Hence it learns because it, or its species, 'must learn'. Its 'goal' entails its existence. Further, its 'goal' is fashioned by the material used in its construction. In a constructed artefact the goal is introduced as part of the cybernetic model.

Now if the learning machine is supposed to imitate an animal, the goal will be an abstraction of certain energetic or material constraints that act upon the real animal. Ashby<sup>16</sup> represents these constraints as a set of 'essential variables' (imaging, for example, the temperature and blood sugar concentration in a real creature) whose values must, in a cybernetic model, be maintained within prescribed limits. On the other hand, if the model simulates the property of learning (without any biological overtones) the goal is somewhat arbitrary and in practically applicable models it is often determined by a desire to optimize process parameters or to recognize specific figures. The important point is that we do not have the liberty in choice of a goal that a cybernetic model suggests. No realisable machine can neglect the constraints and continuities of a material system.

The point is self-evident so far as simple adaptive controllers of the kind considered in 2.2 are concerned. No sane man designs a completely hapless servomechanism, but in this simple case, the machines inhabit a well-defined domain. Their goal, amongst other things, is always obvious. The physics of the situation imposes a partitioning upon the environment that determines a reasonable path whereby the machine can achieve one after another in an hierarchy of subgoals and ultimately satisfy the main objective. To avoid any confusion, let us reaffirm that the information conveyed by a message, coded as a signal sequence, is substantially independent of energy in a macroscopic system. The wise designer, however, gives up the freedom he might exercise and specifies an interaction between the machine and its environment wherein the form of message adheres to the form of energetic or material constraints.

Many of the most useful applications of learning machines (in pattern recognition, the control of really large industrial processes, in management, mechanical translation, man machine symbiosis and in library retrieval), refer to a symbolic or linguistic universe of discourse which constitutes an almost uncharted environment wherein the designer has far greater freedom than either nature or a prudent engineer. Even in the most

favourable cases where the goal is well defined (as it is in pattern recognition and is not in management) there is no guarantee that algorithms exist for achieving this goal or that partitions exist in the set of possibilities.

These issues are, of course, interdependent and Aizerman<sup>17</sup> in his paper, cites a case where a character recognition machine can be designed, with reasonable goal approaching algorithms, because the symbolic environment satisfies a 'compactness hypothesis' which implies the possibility of partitioning (Aizerman's 'compactness hypothesis' reflects a cultural constraint analogous to the energetic constraints of the servomechanism designer. Other principles encountered in this field reflect linguistic, psychological and logical constraints).

It can be shown that learning machines are able to perform abstractions and manipulate some kind of 'concept'<sup>18</sup>. It is argued in 2.4 and 2.5 that they can build non-trivial descriptive hierarchies, associated with 'conceptual' levels and directly related to the subgoals that are searched for in terms of a given level of abstraction. All this is a prerequisite for successful communication with a symbolic environment but it is not enough to ensure effective control. In a given symbolic environment only some 'concepts' can be generalized and, since generalization is the root of coherent inference, only some rather specialized concepts are useful (in the sense that only these yield mechanized inference in place of a haphazard and utterly impracticable search for the solution to a problem). A simple but adequate device is considered by Ullman<sup>123</sup>.

As a conjecture, any knowable symbolic environment is regular enough to admit inference by a machine that embodies the principles that determine its regularity (for, as Greene<sup>19</sup> so often points out, a learning machine cannot be a passive entity that is shaped by its experience. It must, if it learns by inference, have embedded preconceptions about its environment). It may turn out that the embodiment of these principles of regularity, like the evolutionary rules given in 2.7 and 2.8, depends upon a basic property of physical networks. At least one can cite man as the natural existence proof of such a machine.

## 2. Machine Organization

### 2.1. The Organization and the Environment of Learning Machines

Apart from citing a few illustrative cases the first part of the discussion is concerned with information and computation, with the organizations that compute, their input and their output. Later one will examine the physical processes that mediate this organization. Since the minimal criterion for learning is stipulated in 1.3 as a change in the computation that characterizes a behaviour pattern, any learning machine 'learns to compute'. Since this change must be goal directed (again from the criterion given in 1.3) any learning machine also 'learns to control' (even if only in the rather trivial sense that it controls the reinforcement delivered by the experimenter who acts as its environment).

Different kinds of learning machine will be distinguished according to the control procedure they are able to learn, the form of environment in which they learn, and the goals they achieve.

### 2.2. Well Defined and Nearly Stationary Kinds of Environment

Suppose that the environment has the characteristics of a closed (or approximately closed) physical system such as an

industrial process. It is described in terms of variables like temperature and pressure that are well defined, often continuous and always related by process equations. The goal may be to maximize the output of a product. If the system has a stationary behaviour (when it really is closed and its equations really are complete) the learning machine is employed for convenience alone. Its designer could have learned the equations himself and built them into the control mechanism. On the other hand, if the system has a non-stationary behaviour, the controller must adapt its form of computation to maintain the required goal condition. It must be an adaptive controller and may, perhaps, be a learning machine. Andrew<sup>20</sup> has discussed the subject in detail. Although the structure of an adaptive controller is well known, it is cited in Figure 1 for completeness and in order to introduce a symbolism.

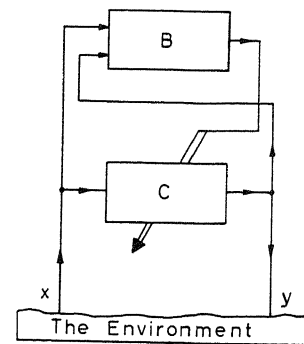


Figure 1

The overall controller  $B$  inspects the input and output state (the outcome or environment state) of a computer  $C$  that controls the environment.  $B$  determines proximity to a given goal as a value of  $\theta(x, y)$ , a descriptor, and changes the function computed by  $C$  (as indicated by the parametric coupling) so that the goal state is approximated. The learning machine  $B, C$ , is a deterministic or stochastic hill climber in which  $C$  obeys a relation of the form

$$y = f_{\phi}(x) \quad (1)$$

where  $x \in X$  is an input state,  $y \in Y$  is an output state,  $f_{\phi}$  is a function  $f \in F$  with an index  $\phi$  adjusted by the output of  $B$  and where  $y$  designates a selection  $y_{t+1}$  if  $x$  is a selection  $x_t$ ,  $t = 1, 2, \dots$ . The box  $B$  obeys a relation

$$\phi = G(x, y) \quad (2)$$

where  $G$  is a strategy chosen to maximize the value of  $\theta(x, y)$ . Minsky and Selfridge<sup>21</sup> have usefully distinguished several forms of hill climbing.

(a) The goal is a unique maximum of a numerical valued descriptor on the machine's parameter space wherein points represent the computed function. The learning machine is a simple optimizer.

(b) The goal is a maximum of several local maxima, in which case the machine must include a 'random' perturbation in order to avoid suboptimal maxima.

(c) The goal is a 'spike in a plain' in which case the strategy of the machine must be 'hill finding' rather than 'hill climbing'. This situation is trivially dealt with by the unpracticable expedient of 'random' search. In fact a very different machine is needed as argued in 2.7 and 2.8, adding a further case to the Minsky and Selfridge categories.

(d) The goal is an optimum sequential decision policy which can be shown to exist in several important cases (in particular when different values of the parameter  $\phi$  determine Markovian subsystems which have the property of generating well-defined values of the stochastic variable to be optimized).

With the possible exception of (a) above (which constitutes a limiting case) the function computed by the adaptive machine is probabilistic in which case the relation of eqn (1) is more conveniently written as:

$$y = x(P^\phi) \quad (3)$$

where

$$P_\phi = \|p_{ij}^{(\phi)}\| = P_\phi(y_j | x_i) \text{ and } i = 1, 2, \dots, n \text{ and}$$

$j = 1, 2, \dots, m$  are indices of  $x_i \in X$  and of  $y_j \in Y$ .

Thus  $C$  selects amongst a set of n.m. conditional probability matrices  $P_\phi$ . If the domain and the range of all  $P_\phi$  is the same (the state sets  $X$  and  $Y$ ) the overall controller may institute a reinforcing strategy. In this case, if  $\theta_0$  is a given value of  $\theta$ , the entire machine computes conditional probabilities.

$$P_\phi(y_j | x_i) = p(y_j, \theta(x, y) \geq \theta_0 | x_i) = p(y_j, \xi | x_i) \quad (4)$$

where the event  $\xi$  occurs if and only if  $\theta \leq \theta_0$ .

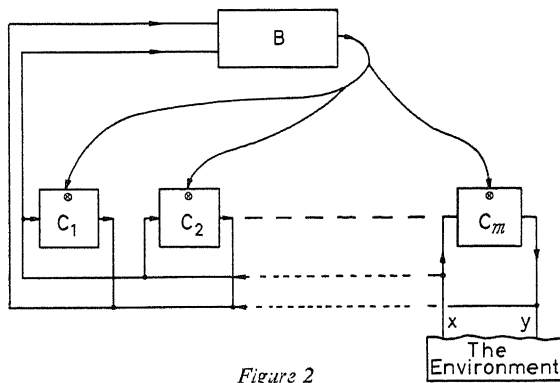


Figure 2

In this case,  $G$  of (2) becomes a function that maximizes  $\theta_0$  so that the system as a whole selects a sequence  $\hat{y} \subset Y$  that, given  $\hat{x} \subset X$ , maximizes the expectation that the value  $\theta(x, y) > \theta_0$  for  $\theta_0$  set as high as possible.

If  $\phi$  is discrete valued (as one assumes) then a useful isomorphic representation of a learning machine is derivable by replacing the parameter variation by a selective process amongst computing machines or subcontrollers labelled in Figure 2 as  $C_1, C_2, \dots, C_m, \phi = 1, 2, \dots, m$  able to compute one of the  $m$  possibly computable functions (the sign  $\otimes$  implies that  $c$  is selected, not that  $c$  receives an input channel, denoted as  $\rightarrow \boxed{C}$ ). Since the inputs and the outputs of the  $c$  are parallel connected, it is unrestrictive to impose the rule that one and only one of them is selected at once. It is, of course, possible to select adaptive subcontrollers, which is tantamount to selecting a selector and that set from which it selects. This image emphasizes the idea of 'selective amplification' first considered by Ashby<sup>22</sup> and the related concept of 'an hierarchy of control'.

The selected subcontroller,  $C_\phi$ , makes selections, when it computes, from  $y \in Y$  on the evidence of a sequence  $\hat{x} \subset X$ . On the other hand,  $B$  selects amongst the  $c \in C$  by assigning a value to the index variable  $\phi$  on the basis of evidence that has been abstracted from the behaviour of the environment when control-

led by its own selection of  $C$ . Hence the selective process is partitioned and the selections made by  $B$  are 'amplified' in the sense that a research director's selective activity is 'amplified' when, instead of selecting from solutions to a set of problems, he selects, instead, from a set of qualified scientists who solve most of these problems on his behalf. Regarded as a computer,  $B$  computes a function  $G$  with domain the values of  $\theta(x, y)$  [an index set of the product set  $(X, Y)$ ] and with range the values of  $\phi$  (an index set of the set  $C$ ). Hence  $G$  is a function of higher logical order than  $F$ , and  $B$  is higher in an hierarchy of control or organization than any subcontroller  $c \in C$ . Indeed, in selecting a subcontroller,  $B$  selects a sign denoting the invariant that this subcontroller maintains. The decisions made by a learning machine that acts as a control mechanism are made about some abstract representation of the environment.

In order to consolidate our ideas, the pioneer work of Uttley and Ashby is briefly considered. Uttley has stressed the process of abstraction and of inference about the existing state of affairs whereas Ashby is mostly concerned with issues of control and stability. These different approaches, however, turn out to be complementary and lead to very similar learning machines.

### 2.3. Learning Machines Derived from Uttley's Conditional Probability Machine

Consider a set of binary variables,  $a^*, b^*, c^*, \dots$  which assume non-zero values if and only if the environment possesses corresponding properties,  $a, b, c, \dots$ . Although there is no limit upon the number,  $M$ , of binary variables that may appear,  $M = 3$  is used for illustration.

The environment can be classified by an abstractive network of AND units, as shown in Figure 3, where a particular output of the network, such as  $a \cdot b$ , is denoted a 'pattern', and a 'pattern' (which is one element in a system of classification) is said to exist in the environment if (in the case cited)  $a^* \cdot b^* = 1$ .

The occurrence of a pattern (except the exhaustive pattern  $a \cdot b \cdot c$ ) does not exclude the presence of one or more other patterns (thus  $a^* \cdot b^* = 1$  does not exclude the possibility of  $a^* \cdot b^* \cdot c^* = 1$  also). However, the network can be made to discriminate up to  $2^M$  states of the binary variables if NOT

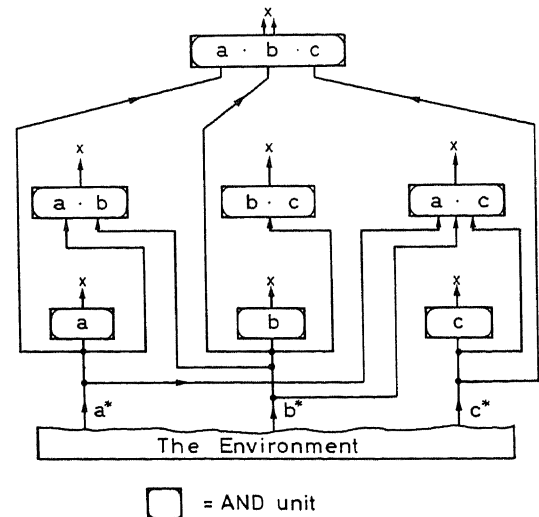


Figure 3

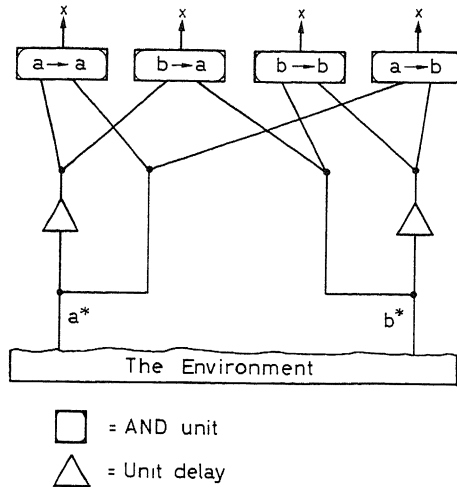


Figure 4

units are also incorporated. The network can also respond to sequences of events if suitable DELAY units are added. In the simplest case, if one and only one of  $a^*$ ,  $b^*$ , and  $c^*$  is non-zero at once, the network of Figure 4 can recognize sequential patterns like 'a after b' and 'b after a'.

If the environment is deterministic, the output of such a network asserts the patterns which exist. On the other hand, if the variables  $a^*$ ,  $b^*$  and  $c^*$  convey an imperfect image of the attributes  $a$ ,  $b$  and  $c$ , its output provides evidence on the basis of which inferences can be made about the patterns that exist. Thus  $a^* = 1$  may provide evidence that  $a \cdot b$  is present even though, on this particular occasion,  $a^* \cdot b^* = 0$  (the underlying assumption is that the environment has stationary statistical characteristics which can be estimated by a machine that has experienced the environment over an appreciable interval).

A conditional probability machine (several have been built) is able to make inferences on the basis of counts and ratios of counts of the number of occasions upon which each output of an abstractive network has previously been energized. The number of previous energizations of output  $r$  is denoted as  $z(r)$  =  $z(r^* = 1)$  when, citing the output  $a \cdot b$  as a typical case, the estimate

$$\frac{z(a \cdot b)}{z(a)} \text{ converges to } p(a \cdot b | a) = p(a \cdot b | a^* = 1) \quad (5)$$

in a stationary environment. Choosing a critical level of  $L$  the machine that embodies these requirements may infer  $a \cdot b$  if, given  $a^* = 1$ ,  $p(a \cdot b | a)$  is greater than  $L$  thus

$$\text{Given } a^* = 1, \text{ infer } a \cdot b \text{ if } p(a \cdot b | a) > L \quad (6)$$

Uttley<sup>13, 14, 15</sup> has built conditional probability machines using logarithmic scale counters and forming ratios by the subtraction of logarithms. He points out that a threshold unit  $r$  with impulse inputs of amplitude  $\eta_i \cdot \beta_i$ , weights  $\alpha_{ir}$  and output  $\eta_r \cdot \beta_r$ , such that  $\eta_r$  is a unit impulse that occurs if and only if

$$\sum_i \alpha_{ir} \cdot \eta_i \cdot \beta_i > \mu_r \quad (7)$$

where  $\mu_r$  is the threshold, can form the basic unit of a conditional probability machine providing each unit includes counters such that

$$\beta_r \text{ is inversely proportional to } z(\eta_r) = z(r)$$

$$\alpha_{ir} \text{ is proportional to } z(\eta_r \cdot \eta_i) \quad (8)$$

These threshold units are arranged in the hierarchical structure of an abstractive network, the output of any unit providing an input term of any unit in which it is conjoined. Thus for units  $a$ ,  $b$ , and  $a \cdot b$  there is the structure of Figure 5.

The value of  $\beta_{ab}$ , given  $a^* = 1$ , is the estimate of  $p(a \cdot b | a)$  and the machine may infer  $a \cdot b$ , given  $a$  if  $\beta_{ab} > L$ . The crucial point of Uttley's construction is that the conditional probability computation depends upon a pair of variables (interpreted as  $\alpha$  and  $\beta$ ). In this connection, Maron<sup>23, 24</sup> has recently demonstrated that a variable threshold element which forms a product from a set of  $m$  weighted inputs can emit an output impulse (designating the assumed truth of an hypothesis) in consonance with an optimum Bayes decision strategy. In the first place it can be shown that an optimum decision strategy for the  $j$ th threshold element given the input evidence,  $\eta_1, \dots, \eta_m, j > m$ , is satisfied by

$$\eta_j = 1 \text{ if and only if } p(\eta_j | \eta_1, \dots, \eta_m) = (\eta_j | \eta_i) > \frac{1}{2} \quad (9)$$

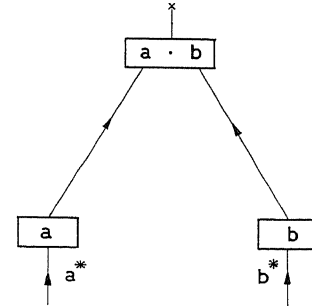


Figure 5

Both the  $\alpha_{ij}$  and the threshold  $\mu_j$  are interpreted as variables dependent upon event counts that converge to  $p$  values (to maintain correspondence with the rest of the discussion, assume that  $\beta_j = 1$  so that  $\eta_j = \beta_j \cdot \eta_j$ ). If these variables are identified as

$$\alpha_{ij} = \frac{p(\bar{\eta}_j) \cdot p(\eta_j | \eta_i)}{p(\eta_j) \cdot p(\bar{\eta}_j | \eta_i)}$$

and

$$\mu_j = \frac{p(\bar{\eta}_j)}{p(\eta_j)}$$

where  $\bar{\eta}_j$  implies the absence of an impulse  $\eta_j$ , so that  $p(\bar{\eta}_j)$  is the probability that  $\eta_j = 0$ , then it is readily shown that (9) will be satisfied if an element has the transfer function

$$\eta_j = 1 \text{ if and only if } \prod_{i=1}^{i=m} \alpha_{ij} \cdot \eta_i > \mu_j \quad (10)$$

Elements of this kind have been simulated by Maron and are physiologically plausible.

Since at least a couple of variables,  $\alpha_{ij}$  and  $\mu_j$  are involved and since a logarithmic measure of the event counts allows the element to decide on the basis of a difference between the threshold and a sum there is a close relationship between this model and Uttley's variable  $\alpha, \beta$ , system.

Regarding Uttley's work, an abstractive network of threshold



units can be regarded as an adaptive filter having an output of patterns, indicated or weighted with a certain amplitude of signal, and an input of binary variables indicating the existence of attributes. After some experience of its environment, this filter, given the index of an attribute, provides a high amplitude output from those channels associated with patterns in which the given attribute has often been conjoined (Thus, given the occurrence of  $a$ , the output of  $a \cdot b$  is high if  $a \cdot b$  has often occurred).

Defining rarity as the negative logarithm of a probability, Uttley considers a conditional rarity machine built in the same manner but from units in which eqn (8) is replaced by

$$\begin{aligned} \beta_r &\text{proportional to } z(\eta_r) \\ \alpha_{ir} &\text{inversely proportional to } z(\eta_r \cdot \eta_i) \end{aligned} \quad (11)$$

which adapts to become a filter that given the occurrence of an attribute provides a high amplitude output from channels associated with patterns in which the given attribute is rarely conjoined. (Thus, if  $a \cdot b$  often occurs and  $a \cdot c$  rarely occurs then, given  $a$ , the output of  $a \cdot c$  is of high amplitude and the output of  $a \cdot b$  is of low amplitude).

It has been assumed that an abstractive network is specified (so that the relevant patterns in the environment have already been determined). If this is not the case (and commonly it is not) it would be possible to build a complete network, with  $2^M$  outputs. Providing that all the relevant attributes have been indexed by binary variables, the relevant patterns will exist in its output. However, the number of elements required to build this network becomes gigantic for large values of  $M$  and the construction would be impractical. Further, in a control mechanism, the relevant patterns which constitute the evidence for a decision are normally restricted. Hence it appears necessary to start off with uncommitted elements which the machine connects together depending upon its experience of the environment by some process of maturation which might be called 'learning an abstraction'. The first possibility is that connections develop between coincidentally stimulated units  $a$  and  $b$  and a unit that comes to index their conjunction  $a \cdot b$ . Uttley discards this hypothesis because it entails the existence of a unit labelled  $a \cdot b$  to recognize the  $a$  and  $b$  conjunction in the first place. The alternative hypothesis is an over connected network of units which differentiates and becomes more selective as a result of its experience. The organization of a conditional rarity machine can be shown to develop in this manner.

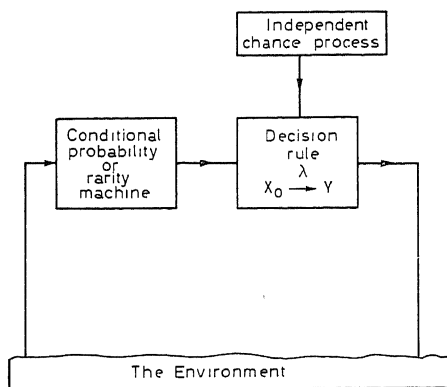


Figure 6

Suppose there is a conditional probability (or a conditional rarity) machine delivering indices of the conditional probabilities of several states  $x_v \in X_0$  denoted as  $p(x_v | x_i)$ , given any  $x_i \in X - X_0$  where  $X_0 \in X$  is a particular subset of states entailed by a control procedure with a decision rule  $\lambda$ , from  $X_0$  into  $Y$ . This machine, combined with an apparatus that embodies  $\lambda$ , is a control mechanism.

The environment can only be specified statistically. Hence the decision rule  $\lambda$  is embodied in a probabilistic device that selects  $y \in Y$  as a function of an independent chance-like process that is biased by  $\varepsilon \|p(x_v | x_i)\|$  as in Figure 6. Thus the system's behaviour is expressed by

$$y = x(P)$$

$$\text{for } x \in X - X_0 \text{ where } P = \|p(y_j | x_i)\| = \|\lambda[p(x_j | x_i)]\| \quad (12)$$

which is consonant with eqn (3).

The formulation is essentially unchanged if the output selects amongst sequences of motor activities, rather than distinct response states.

To complete the picture, an adaptive reinforcing procedure must be introduced. Several mechanisms have been proposed by Uttley and Andrew, of which the most convenient is to adjoin a further state,  $\xi$ , which exists if and only if  $\theta > \theta_0$  as in (4), to each conjunction in the abstractive network. Thus the machine computes conditional terms, which are 'conditional upon reinforcement', yielding a set of values

$$p(y \cdot \xi | x)$$

as in (4) and comment that the abstractive system shown in Figure 7(a) is characterized, at a particular instant, by a matrix

$$P_\phi = \|p(y \cdot \xi | x)\|$$

where  $\phi$  is the adaptive parameter.

Variation in the parameter  $\phi$  will adequately describe the maturation as well as the adaptation of a probabilistic machine

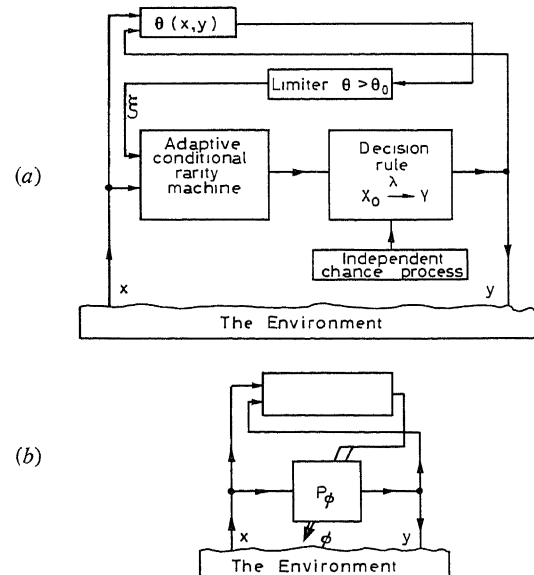


Figure 7. In practical systems direct  $\theta$  reinforcement is used, or, alternatively,  $\theta_0$  is changed each finite interval by an adjustment  $+\Delta\theta$  if the mean value of  $d\theta/dt$  over an interval  $\tau$  exceeds 0, or  $-\Delta\theta$ , if not (a) and (b) are functionally identical

providing that it is always presented with the same statistically stationary environment. When the environment is non-stationary however, the  $p$  values are in continual flux. If the environment is characterized by several distributions  $\Pi_1, \Pi_2, \dots, \Pi_L$  there may be up to  $L$  different successful adaptations and the machine will have to re-adapt whenever  $\Pi_r$ ,  $r = 1, 2, \dots, L$  is changed. Thus, having adapted to  $\Pi_r$  by forming  $P_{\phi r}$  the machine may be presented with  $\Pi_{r+1}$  when it must start to adapt once again. As a result of achieving the successful adaptation  $P_{\phi r+1}$  it loses all trace of  $P_{\phi r}$  and is at no particular advantage if  $\Pi_r$  occurs subsequently. What is needed (and what may readily be provided) is an attention mechanism  $\Omega$  as in Figure 8 selecting amongst differently matured networks according to evidence from the immediate behaviour of the system. Each box  $M$  in Figure 8 represents a complete machine of the form II in Figure 7 selected by  $\Omega$  according to the convention of Figure 2.

Other probabilistic learning machines (Steinbuch's<sup>25</sup> learning matrices, a system due to Ratz and Thomas<sup>116</sup>, and the 'Eucrates' systems devised by Bailey, McKinnon Wood and Pask<sup>26</sup>) closely resemble the original formulation of Uttley's. They differ chiefly in emphasis (in the 'Eucrates' system, for

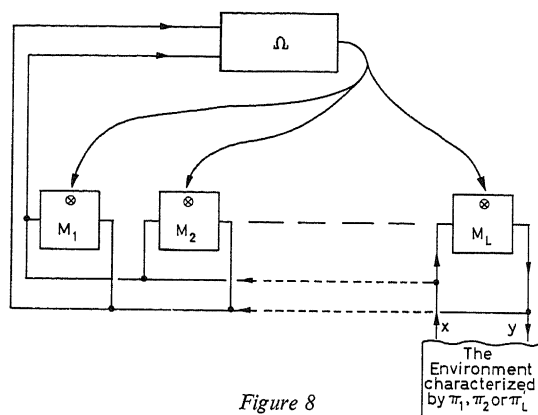


Figure 8

example, the idea of an attention mechanism is emphasized) or in identification (they are not, as the Uttley model is, wedded to a neurological hypothesis).

#### 2.4. Other Learning Machines Built from Networks of Elements

There are many adaptive networks in which the coupling weights  $\alpha_{ij}$  between the elementary units are varied as a function of the activity of the system and an externally manipulated reinforcement variable. Taylor<sup>27, 124</sup> for example, has an artefact made from linear elements (the impulse rate at the output is proportional to the weighted sum of the impulses arriving at several inputs). Widrow<sup>28</sup>, Willis<sup>29</sup> and others have built networks of threshold devices (which satisfy eqn (7) with  $\beta$  a constant) and Babcock<sup>30</sup> constructed a special purpose computer in which the elementary units can be variously programmed. All of these systems can be induced, by suitable training, to respond in a given fashion to patterns of input stimuli. Probably the largest body of data in this field is due to Rosenblatt<sup>31</sup> who has fabricated many different adaptive pattern recognition machines called 'Perceptrons'. A typical Perceptron is shown in Figure 9. The 'response' units are threshold units with constant weights. The 'association' units are threshold devices with variable weights  $\alpha_{ij}$ , the change  $\Delta\alpha_{ij}$  depending upon the correlated activity

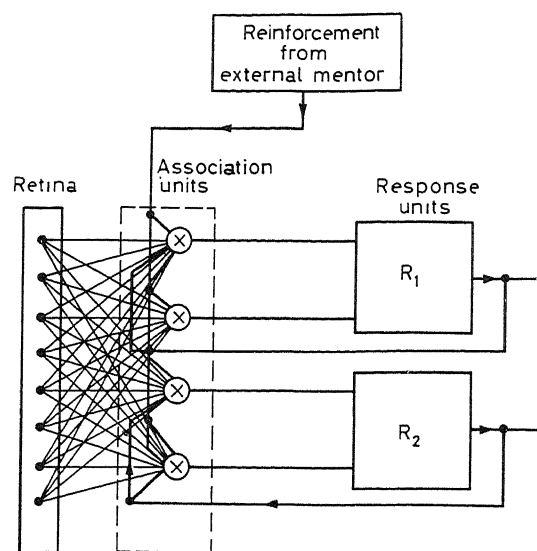


Figure 9

appearing at the input and output of the  $j$ th element and upon the approval of the external mentor. Rosenblatt's original description allowed for 'random' connectivity. More specialized versions of this system have properties that vary from a passive logical filter, able to abstract a given attribute of its input to self-oscillatory systems within a back coupled mesh. Between these limits are networks that act like conditional probability mechanisms. The idea of 'random' connectivity, which appears rather often in the literature, is more fruitfully reinterpreted as a modicum of overconnection in the sense cited in 2.3 that is removed as a result of adaptation. In the first place the network is neither 'random' for the reasons cited in 1.8, nor is it 'random' in the sense of being unrestrictedly variable. As Papert<sup>32</sup> points out, any  $m$  input network which was unrestrictedly variable (in the sense of being able to compute all  $2^{2^m}$  functions of  $m$  binary input variables) would be unrealizable for sensibly large values of  $m$  since no training strategy could reach the desired adaptation within an acceptable time. Fortunately, the issue does not arise in practice for many special limitations are built into the specification; for example, threshold elements can only compute linear discriminating functions (if the  $\alpha_{ij}$  are coordinates of a state space the values of the  $\alpha_{ij}$  determine a linearly discriminating plane  $\sum \alpha_{ij} \eta_i = \mu_j$  and  $\eta_i = 1$  if the input state is on one side of this plane and  $\eta_j = 0$  if it is on the other). This case is examined by Papert<sup>32</sup>, Singleton<sup>33</sup> and Scott Cameron<sup>34</sup> who show that the percentage of Boolean functions that are linearly discriminating decreases rapidly with increasing value of  $m$ . Ivankhenko<sup>35</sup> considers similar problems in the case of a perceptron.

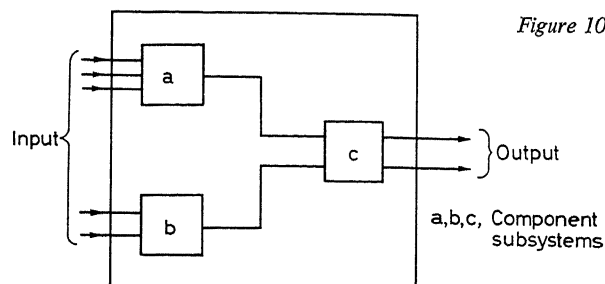


Figure 10



It may turn out that the class of linear discriminating functions, although tractable in the sense that it is a small enough domain for an adaptation strategy to work in a reasonable time, is a rather inept choice. Many other classes of Boolean function would be acceptable. Willis<sup>36</sup>, for example, proposes networks restricted to the computation of disjunctively decomposable functions of their input and it can be shown that such a network is separable into components, as indicated in Figure 10. It is important to avoid confusion between this separation of components and the hierarchical organizations considered in 2.5.

### 2.5. Ashby's Homeostat and More Elaborate Hierarchically Organized Controllers

The Homeostat is shown in Figure 11 in the form described by Ashby<sup>37</sup> (a more versatile apparatus has recently been developed by Haire, Harouless, Miller and Williams<sup>38</sup>) and a homeostat with memory has been constructed by Chichinadze<sup>119</sup>.

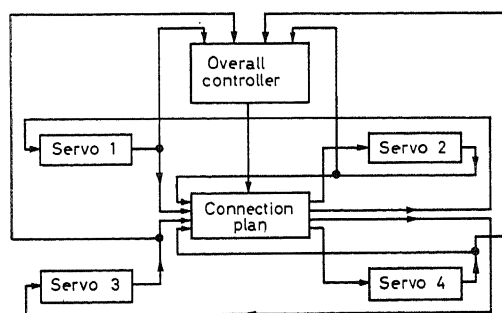


Figure 11

The original Homeostat consisted of four null output positional servomechanisms. The positional output of any one can be applied (depending on a plan of interconnection) to the input of any other unit. Each positional output potentiometer is provided with limit indicators and, in the original demonstration, was identified with one essential variable of an organism. Ashby used the apparatus to demonstrate 'ultrastability' which is a generalized version of adaptive control. Suppose that the homeostat has a plan  $P_1$  of interconnection and is given an input (from some external environment) that perturbs several of the units. Now, given  $P_1$ , the system may or may not be stable. If not, at least one of the essential variable potentiometers indicates a limit value (which is taken to contravene the stability condition for the system). In this case, an overall controller (equivalent to  $\Omega$ ) is required to select another plan of connection between the units, say  $P_2$ . If the system is stable, given  $P_2$  and a typical environmental input to perturb it, no further change occurs. If not, the overall controller makes a different selection.

In the homeostat, the overall controller strategy is the most generalized (and also the least provident) conceivable. The machine aims merely for stability,  $P_1$  and  $P_2$  being selected in an arbitrary fashion and independently of the state of the system. An adaptive controller  $B, C$ , is a specialized version of this paradigm. However, note that the boxes  $c \in C$  do not correspond to the physically discriminable units in a homeostat. They correspond, instead, to subsystems created by an interactive coupling specified by  $P_1$  or  $P_2$ , between these units. The computing

machines are well defined organizations, but they are not necessarily well localized physical entities. This comment has marginal relevance in connection with simple homeostasis but it becomes important when considering extensive hierarchies of control as in Figure 12.

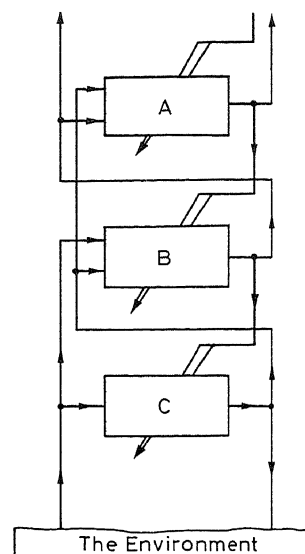


Figure 12

MacKay<sup>39, 40</sup> was the first to consider large, hierarchical, assemblies of computing systems. Using the argument of 2.2. (which is suggested by MacKay) a mechanism  $A$  may be regarded as selecting amongst signs that denote the invariants maintained by subcontrollers  $b \in B$ , the  $b \in B$  may be viewed as selecting amongst signs associated with the  $c \in C$ . Hence the entire learning machine is capable of abstraction (in particular of abstracting a goal or homeostatic criterion). Andrea<sup>41</sup> discusses the issue in detail and Mesarovic and Banerji<sup>117</sup> have developed a theory of a self-organizing system based mainly on the idea of an hierarchy of control.

If the goal is well determined and unchanging, the structure has little interest. There are, however, many environments to be controlled in which the goal is nebulous, for example, if the control mechanism is used in management it will probably be asked to maintain 'happiness' and 'productivity'. Now even if one presupposes that a machine construction for 'happiness' or 'productivity' maintenance can be upheld, it is still necessary to admit that these properties are differently interpreted upon different occasions, and that the level as well as the form of abstraction involved in their expression is variable. If the machine were less flexible, control would be impossible, because the required invariance is specified with reference to a parameter which is given different ostensive definitions upon different occasions.

A very similar dilemma is engendered when the environment to be controlled is made up from incomparable parts. In this case, the control problem is solved by a machine that contains an attention mechanism as given in 2.3. In order to achieve a given goal, this machine must 'attend' to different inputs upon different occasions and 'contemplate' different response alternatives.

With reference to it can be said that any machine which

which includes an hierarchy of control, or of goals, can, in certain circumstances, be said to contain an hierarchy of metalanguages and that any machine able to build such an hierarchy of control can, in the same circumstances, be said to 'construct' metalanguages. Manifestly, such a machine can build abstractions and, given the provisions of 1.5 it may be said to use 'concepts'<sup>42, 43</sup>.

## 2.6. A Self-organizing System, Localized, and Unlocalized Automata

Before considering the issues of attention and goal variation in earnest, let us examine one, slightly more tractable, system. This is a special case of a so called 'self-organizing system' (it is commonly agreed that the term 'self-organizing system' is unfortunate since the entity concerned is really a sequence of systems). Now according to von Foerster<sup>44, 45, 46</sup> a subsystem  $J_r$  (one member of this sequence) is a self-organizing system if the rate of change of its redundancy,  $R$ , is positive

$$dR(J_r)/dt > 0 \quad (13)$$

The special case to be considered is a model that represents the data processing entity which is referred to as 'a man' when one speaks about a 'man' learning or performing a realistic skill.

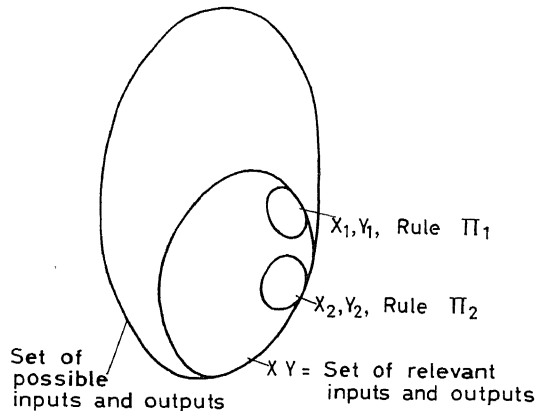


Figure 13

There is plenty of evidence that a conscious 'man', regarded as a data processor or computer, has the characteristic that in order to survive he must maintain a certain rate of adaptation (the environment must provide sufficient novelty for him to learn about)<sup>44</sup>. Another undeniable fact about a 'man' is that he is, over any finite interval, a finite automaton with a well defined input called his 'field of attention' (and, incidentally, a well defined output of responses that in some way act upon his 'field of attention'). In so far as the training routine or the actual skill occupy his attention, the field of attention, with relevant input states  $X_r$  and output states  $Y_r$  is coupled to relevant stimuli and responses as suggested in Figure 13. This condition would not be satisfied if the man were bored or inattentive. In order to satisfy the basic requirement that a certain rate of adaptation is maintained, the environment must be capable of offering subsets of stimuli and responses ( $X_r, Y_r$ ) that have a statistical relationship  $\Pi_r$  which the man can learn. Where  $(X_r, Y_r) \subset X, Y$  is the product the set of relevant stimuli and response states.

This primitive image of man can be modelled with a set of the conditional probability machines given 2.3 and character-

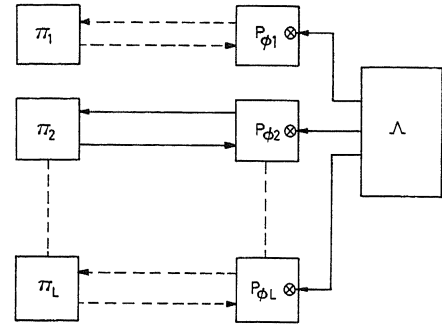


Figure 14. Selection of  $\pi_2, P_{\phi 2}$ , is shown

ized by matrices  $P_{\phi r}$  as in eqn (3) and having input states  $X_r$  and output states  $Y_r$ . The environment is characterized by a set of matrices  $\Pi_r$  (which may also be embodied in a machine as suggested in Figure 14). It will be convenient to assume that  $P_{\phi 0r}$  has equal  $p$  entries. Calling the  $r$ th coupled subsystem consisting of 'man' and his immediate environment  $J_r$ , it is obvious that the adaptation rate condition is equivalent to eqn (13). In this model one can insist that  $J_r$  will maximize its chance of survival, say by computing  $R(J_r)$  as an index of adaptation and, from eqn (4), making

$$\theta_r = R(J_r) = 1 - \frac{H(J_r)}{H(J_r)_{\max}} = 1 - \frac{H(X_r, Y_r)}{H(X_r, Y_r)_{\max}} \quad (14)$$

and since  $H(X_r, Y_r)_{\max}$  is invariant for a given coupling, one has  $\theta_r \approx -H(X_r, Y_r)$ . Thus, if eqn (14) is substituted in eqn (4) and if the values of the adaptive parameter  $\phi$  converge,  $\phi_0 \rightarrow \phi_1 \rightarrow \dots T$  one obtains

$$d\theta_r/dt \approx dR(J_r)/dt > 0 \quad (15)$$

for  $\phi \neq T$ . Wattanabe<sup>47</sup>, for example, has considered the dynamics of systems like this which *must* adapt and points out that learning behaviour as well as learning models satisfy

$$R(J_r) | t \approx -H(J_r) | t = a(t - t_0)e^{-bt}$$

where  $a$  and  $b$  are positive constants and where  $t > t_0 \geq 0$  is the instant of observation.  $J_r$ , however, is obviously an unstable or ephemeral self-organising system. Since the reinforcement variable,  $\theta$ , is increased by any increase in the regularity of  $J_r$  behaviour, the system will approach a stable (though possibly dynamic) equilibrium characterized by

$$P_{Tr} = \text{stochastic inverse } (\Pi_r) \quad (16)$$

However, if eqn (16) pertains,  $J_r$  is fully adapted, hence, at this point,

$$d\theta_r/dt \approx dR(J_r)/dt = 0 \quad (17)$$

contravening eqn (13) so that  $J_r$  is no longer a self-organizing system (it can be demonstrated that although a 'forgetting' facility can delay the instant at which eqn (17) applies, 'forgetting' does not avoid the ultimate instability). When this occurs, a *real* 'man' will 'change his attention'. To comprehend this fact, any model (cf. Figures 14 and 17) must include the transformation, say  $\lambda$ , embodied in a mechanism  $\Lambda$  with domain a set of  $M$  or more subsystems  $J_r, J_r \in J$ , and with range  $J$ . Let  $\Lambda$  receive an instruction, whenever eqn (17) is true, to select  $\lambda(J_r) = J_{r+1}$

which consists of a conditional probability machine characterized by  $P_{\phi, r+1}$  coupled to an environment characterized by  $\Pi_{r+1}$ . For an initial selection of subsystem  $J_0$  and for  $M$  iterations of this process the minimal constraints

$$\lambda(J_r) \in J \text{ and } \lambda(J_r) \neq J_r$$

yield a sequence

$$J^* = \bigcup_{r=1}^{r=M} [\lambda^r(J_0)] \quad (18)$$

such that  $dR(J_r)/dt > 0$  for all  $J_r \in J^* \subseteq J$ . Thus  $J^*$  is a self-organizing system.

This model serves as a very primitive image of 'man' providing (a) that the adaptation takes place with reference to some goal (like performing a given skill), not just arbitrarily (increasing  $R$  is tantamount to performing any skill), and (b) that although, in the limiting case,  $\lambda$  may act when eqn (17) is true, it is also possible to replace this instruction with an internally or externally produced 'attention directing' signal.

A more basic and interesting distinction between this model and reality is that, in fact, the set  $J$  is not denumerable.  $J$  is defined (and consequently renders the disjunction of eqn (18) legitimate) in terms of some property of  $J^*$  which leads to assertions like 'these systems all represent the behaviour of a man', or 'of an animal' (which, since they refer to the systems rather than the behaviours, are metalinguistic statements).

Typically, this situation comes about when there is a reproductive process capable of creating and replicating machines characterized by  $P_{\phi r}$  which survive if and only if, when coupled to form systems like  $J_r$ ,  $d\theta/dt > 0$ . In this case, if a sequence of  $M$  systems does survive, the transformation  $\lambda$  is interpretable as a rule of evolution that, in a specified environment, gives rise to  $J^*$ , and the members  $J_r$  of  $J^*$  are characterized by the property that they are produced by this evolutionary rule.

The point leads us to a different, 'evolutionary', kind of learning machine. The previous learning machines stem from the familiar model of a localized and finite automaton (a finite automaton with well defined input states and output states). The model for evolutionary systems is a population of reproducing automata.

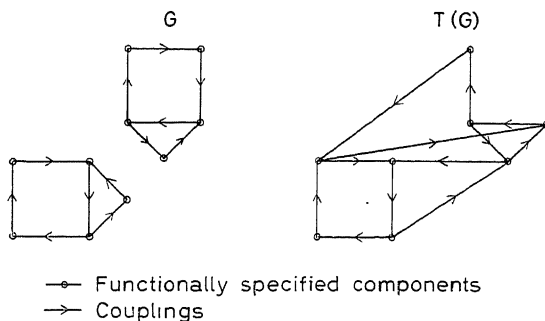


Figure 15

Suitable constructions have been developed and discussed by Burke<sup>48</sup>, von Neumann<sup>49</sup>, and Loeftgren<sup>50, 51</sup>. In Loeftgren's development of von Neumann's work, it is pointed out that a dual representation of these automata is commonly and perhaps necessarily adopted.

(a) As a developing connection graph (with which there is an associated state graph). This method has been adopted by

Rashevsky<sup>52</sup>, Rosen<sup>53</sup>, and others. A transformation,  $T$  in Figure 15, maps graph  $G$  into graph  $T(G)$ . Often enough  $T$  cannot be readily expressed in an analytic form.

(b) As a configuration on a 'tessellation' (a 'tessellation' is an infinite plane of cells,  $i, j$ , each of which can assume one of a finite number of states  $u \in U_i, u \in U_j$ , amongst which, for each value of  $i$ , at least one state is a null state denoted as  $u_0 \in U_i$ ). Entry into a null state is interpreted as the disappearance of some aspect of an automaton, which is a configuration of the states of certain neighbouring cells, and any transition from the null state implies the creation of some feature of an automaton. The transition rule,  $Z$ , depends, for the  $i$ th cell, upon  $u \in U_i$  and the state of neighbouring cells. Hence  $Z$  is a mapping which if  $i$  and  $j$  are neighbours is of the form  $Z; (U_i, U_j) \rightarrow (U_i)$ .

If an automaton is localized its (a) image is a structure of functionally defined components and its (b) image is a connected configuration of other-than-null states. If an automaton is unlocalized, in particular if it 'evolves', its (a) image is often difficult to interpret because of an ingrained confusion between organizations and objects, whereas its (b) image is a replicating population of distinct configurations.

It is worth noticing that:

(1) In his discussion of 'error correction', Loeftgren shows that in a real universe, where an automaton receives permanent structural perturbations as well as irrelevant inputs, no localized automaton can have an infinite life span but that some unlocalized automata can have an infinite life span. This distinction demarcates the set of evolutionary learning systems. It is important, however, to avoid confusion between this and the well-known distinction of minimal automata and automata that are deliberately built with structural redundancy in order to effect error correction.

(2) As emphasized in 1.2, the simulation of a learning machine entails the simulation of its normal environment. If learning is interpreted as the reproduction of an organization (which is, for example, the view adopted by Wiener) the environment can be said to mediate the reproductive process. In fact, this interpretation (or some equivalent) is necessary in order to avoid Rosen's<sup>54</sup> paradox (which is a special case of Russell's paradox), but the term 'environment' implies an internal 'environment' such as a brain or some other physical embodiment in which the evolutionary process can be realized. It can be called, unspecifically, the medium in which evolution takes place<sup>55</sup>.

Because of the confusion between organizations and objects it is often difficult to specify the medium in the (a) image, but the (b) image shows a population in (or an organization that is a property of) the most abstract possible representation of a medium, namely a tessellation. This is a straightforward picture and, for most purposes, is preferable.

Discussion of abstract systems that learn, in the evolutionary sense, is always related to the dual pictures of (a) and (b) but some contributors to this field, such as Masano Toda<sup>56</sup>, Heinz von Foerster<sup>61</sup>, Caianiello<sup>58</sup>, and Pask<sup>59, 63</sup>, have been differently motivated and oriented. A far less abstract version of a medium is used, akin to a close coupled, but real environment. However, it remains true that the evolving organization is a property of the medium rather than something that exists in its own right and a couple of points are worth emphasis. In the first place, the evolutionary rule,  $Z$ , is a property of the other than minimal specification of the medium thus illustrating the material

dependence cited in 1.5. Next, the medium in which learning systems can evolve is highly constrained, but the constraints do not refer directly to the computations manifest in the learning behaviour. They exist in order to ensure that evolution of any system can take place.

## 2.7. Different Realizations of Learning Machines in Evolving Systems

Selfridge<sup>61</sup> has devised a hierarchically structured programme called 'Pandemonium' that can be used either for recognition or for control. The 'Demons' in a Pandemonium are computing routines and are isomorphic with possibly adaptive subcontrollers, as shown in Figure 16. The lowest order demons, in contact with the environment, recognize or control predetermined features, indicating their achievement by signals conveyed to higher order demons that assign weights to these signals and transmit a linearly weighted evaluation of the lowest demons, performance to an overall controller demon which is criticized by some external mentor.

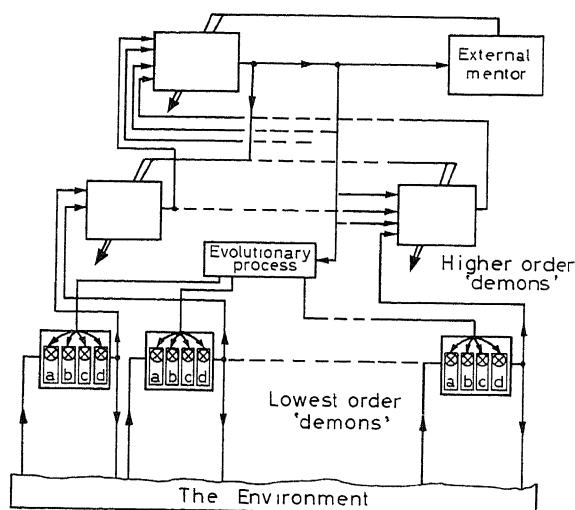


Figure 16

The system resembles a Perceptron in so far as the middle demons perform a linear weighting and in so far as there is some external mentor. It differs a great deal in respect to the lowest demons, which in Pandemonium are highly specific feature recognizers or subcontrollers. A more important difference lies, however, in the development of the system, which is not revealed by the instantaneous picture. Whereas adaptation in a perceptron entails only an adjustment of weights, the demons in a Pandemonium must be capable of evolution. The designer is supposed to be prudent in his choice of features, but not omniscient. Consequently, subcontrollers that meet with common disapproval, evidenced by uniformly low weight assignments, must be discarded and others created to take their place. If a subcontroller in the organization of a Pandemonium is identified with a species of automata, the organization could be realized by a fairly simple evolutionary process. Members of a population of automata compete for some commodity (food or money) that is available in short supply and which allows them to persist and to reproduce (this tacitly assumes a cost, in terms of food or money, for maintaining the fabric of an automaton and a similar cost for its replication). The supply of food or money is sufficient to maintain, on average, a number,  $n$ , of automata, which (depend-

ing somewhat upon the distribution of automata in different species) can sustain the activity of the required number of subcontrollers. Since more than  $n$  automata are created there is competition and some automata fail to survive (indeed, those that survive will be members of species favoured by the higher order components that act, presumably, upon evolutionary rules).

Given some kind of variation, there will be a selection of favoured variants. The issue is, what principle of selection is to be used? Purely chance variation would be pointless, for the set of possibilities is too large, but two mechanisms are possible, namely:

(1) Recombination. Instead of discarding an unpopular demon and creating a novel demon, the elements of a partially successful demon are recombined in a fresh arrangement. Hence, amongst other things, the adaptations of the original demon are not completely lost.

(2) Co-operative interaction whereby the original demons are coupled into co-operative entities which are reproduced, by some different and possibly evolved mechanism, as a whole.

Although a model of this kind is easy to describe and parts of it have been successfully simulated, there is a tendency to consider media more like brains than the environments of simple automata. Pringle<sup>62</sup>, suggested that hierarchies of control could evolve in a large multiple mode oscillator due to the non-linear coupling between stable oscillatory modes. Such a system is conveniently realized by a large network of self-exciting threshold elements, with malleable coupling. The existence of a stable mode imprints a pattern upon this medium which, in turn, induces the stable mode or filters its components from uncorrelated excitation (in a stable system there will be a many to many correspondence between patterns and dynamic components). The existence of a stable mode acts as an informational 'metabolism' that maintains the constraints which are required in order that the network shall compute a specified function of its external stimulation (hence modes of oscillation are, in this case, the subcontrollers of demons). Networks like Beurle's simulation<sup>9, 10</sup>, mentioned in 1.2 are admirable media for this purpose and Beurle has examined the self-excited as well as transmissive characteristics of this system. Amongst others, Farley and Clark<sup>63</sup> have simulated very large assemblages of somewhat simplified non-linear elements. Taylor's paper, at the present meeting, deals with this area.

From a theoretical viewpoint, Wiener<sup>64</sup> has recently proposed a potentially quantitative theory of learning and has tentatively applied it to data from cortical activity. As a generalization of our previous comments about oscillatory modes, replication (which is viewed as the basic mechanism of learning) is represented by the distortion free propagation of a wave of activity in a medium of non-linear elements. (The comments on filtering are generalized in terms of Wiener filters.) Further development of this work should lead to a precise and comprehensive image of the evolutionary mechanism that underlies learning<sup>65</sup>.

## 2.8. Problem Solving, Communication and Strategy in Linguistic Environments

As suggested in 1.5, there are many situations in which the control engineer is required to build learning machines that interact with a symbolic environment. Often these situations are obvious enough; for example, the related issues of language translation and data retrieval certainly entail learning (at any

rate in large systems) and it seems unlikely that the learning of sign sequence correlations is sufficient (the machine must learn the language concerned). The same is true for the whole of 'artificial intelligence'. Other situations are not so obviously symbolic in character. Tustin<sup>66</sup> has demonstrated that many facets of the economy may be predicted and perhaps controlled within a model that is derived from Laplace transformations of economic data. As a result, the design of a control mechanism appears to involve an application of conventional methods. However, there are discontinuities, due to invention, for example, which prove intractable within the framework of Tustin's model. One can, of course, neglect these or, failing that, one can interpolate a guess, but it seems at least possible that a machine capable of symbolic interaction might guess solutions to these problems, better than oneself, but by using ones own methods which are (by hypothesis) to discern relevant signs in an underlying cultural language and to interpret them as indices of change (this, incidentally, will involve feedback perturbations of the controlled system and a machine capability for learning as well as manipulating the language).

Another situation involving symbolic interaction is man machine symbiosis, for example, in problem solving (the term 'symbiosis' is used to distinguish the present relation between the man and the machine from the more familiar relation in which a machine is used as a problem solving tool). In symbiosis, the aim is man machine co-operation. The machine should suggest hypotheses and proofs to the man, as well as carrying out his instructions. In order to do this, it must develop a common language (so that the man machine interaction has the logical calibre of a conversation). As a result, in a genuinely symbiotic system, it would be impossible to tell whether the man or the machine was responsible for the solutions achieved.

Very little work has been done, so far, in symbiotic problem solving but some results are available for comparable symbiotic systems used for training (the control of learning) and mental testing (stabilizing the man machine interaction to maximize the information available regarding the values of specified parameters).

A few specific mechanisms are now discussed in an attempt to illustrate some of the chief difficulties of the field.

In order to talk about signs, it is first necessary to specify an object language  $L_0$  (which is used by the system concerned) in terms of an observer's metalanguage  $L^*$ . As Cherry<sup>67</sup> has pointed out, this expedient is a formal prerequisite for any kind of experimentation (but it is tacitly assumed and taken for granted when the experimental technique is well known and the  $L^*$  interpretation of  $L_0$  signs is invariant). If one also wishes to assert that an organism or a machine manipulates signs (in particular, that it learns about signs or acquires concepts) one must either:

(a) Be in a position to embed symbolic representations within the machine (trivially by building them into it and not entirely trivially by the ham-fisted reinforcement of an adaptive network so that it recognizes the symbols determined by its external mentor), or

(b) Understand a common environment (by knowing 'essential variables', for example, so that, in the absence of reinforcement it can be taken for granted that the organism or machine will learn to use signs denoting the critical value of these 'essential variables' if it survives. Failing this, it would encounter the critical condition rather than a sign for it).

Returning again to 1.5, the fact that not all abstractions can be generalized is tantamount to the fact that not all signs act as symbols for a given machine and its environment. To cite a picturesque but defensible analogy, the population of automata in an evolutionary learning system must (if they interact co-operatively as in 2.7) possess an 'internal' language with symbols of its own. Generalization is possible if this 'internal' language can be translated into the external language and *vice versa*.

Having established  $L_0$  and represented it in  $L^*$  a control mechanism can be regarded as a problem solving device. The perturbations to which this controller responds correctively pose problems (certain equivalence classes of  $x \in X$  denote problems and certain  $y \in Y$  that index the controller's asymptotic response denote solutions). The controller becomes a problem solver (an adaptive controller becomes able to solve a particular problem) and the functions computed by the fully adapted machine specify a decision rule. A largely equivalent formulation images the machine as a game player when certain  $y \in Y$  are its moves, certain  $x \in X$  are the moves of another player,  $\theta(x, y)$  is the payoff function of the game over the outcome set  $X, Y$ , and the computed function determines strategies,  $\hat{y}$  (as described, the game is played in its extensive form).

An abstractive or hierarchical learning machine is capable of quite elaborate problem solving and game playing since it can learn to solve a class of problems by constructing descriptions of problem solution. An hierarchical structure can embody as many machine metalanguages  $L_1, L_2, \dots$  (describing and manipulating  $L_0$ ) as there are levels of control, and evolutionary systems are able to build  $L_r, r = 1, 2, \dots$  (which are, in some sense, interpretable as conceptual languages).

The issue of 1.5 reappears, however, namely 'in what sense are these levels of control conceptual levels?' Providing there are no overtones of mentation, any abstractive learning machine can have and some can build 'concepts', but in order to remove the inverted commas, it is necessary to consider the discourse between this machine and other machines or other men.

A pair or more of learning machines can interact at several levels of discourse,  $L_0, L_1, \dots$  either through couplings established between different levels of control or, due to similarity of construction, by inferring the  $L_1, L_2, \dots$  implication of expressions in  $L_0$ . George<sup>68, 69</sup> has experimented with multi-machine computer simulations and has examined the linguistic requirements entailed in the successful conduct of non-zero sum, partially co-operative, games and certain sequential games. The entire experiment is explicated, of course, in  $L^*$ .

On the other hand, in connection with 'artificial intelligence', it is not only necessary to describe the problem solving in terms of  $L^*$ , but also to interact, personally, with the machine.

A typical man machine system is proposed by Edwards<sup>70</sup>. The job of constructing a Bayes estimate of the probability of an hypothesis, conditional upon certain data

$$p(\text{Hyp} | \text{Data}) = \frac{p(\text{Data} | \text{Hyp}) \cdot p(\text{Hyp})}{p(\text{Data})}$$

is ideally divided between a man and a machine. There is evidence to suggest that a man can be trained to accurately determine  $p(\text{Data} | \text{Hyp})$  which a machine cannot readily do (especially if much of the data is irrelevant or if the hypothesis is tentative) providing he receives a guide from present values of  $p(\text{Hyp} | \text{Data})$ . On the other hand a man is a notoriously in-

accurate calculator and in Edward's system, the Bayes' calculation is conducted by a machine.

The extreme case of man machine interaction, conversational man machine symbiosis, is exemplified by the adaptive teaching system in Figure 17 (these systems have been used to instruct a number of industrial skills including perceptual motor skills, like card punching and intellectual skills like fault detection on the basis of imperfect evidence)<sup>71-74</sup>.

Broadly, the machine acts like a personal instructor who changes the training routine to suit an image of the student, derived from continual observation, by posing novel problems

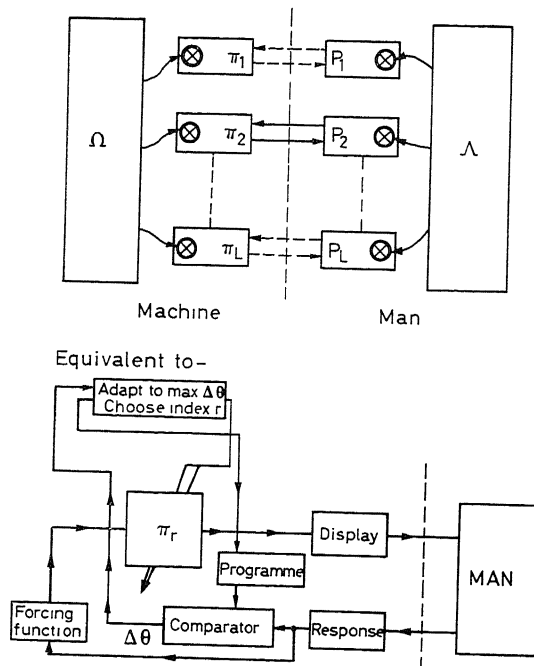


Figure 17

when the student is bored and simplifying the material or co-operating if the displayed data is too difficult. Precisely, the system is a stochastic optimizer based upon the model given in 2.6. The machine selects subenvironments, characterized by the different matrices  $\Pi_r$ , with an operation  $\Omega$  analogous to the operation  $\lambda$  (in the 'student' part of the model). Adjoining the plausible condition that each of the adaptive 'automata' (that momentarily represent the student) must receive intelligible problems to solve, and also interpreting  $\theta_r$  as an index of relevant regularity derived from a success measure in such a way that  $\Delta\theta_r$  is an index of learning rate in the  $r$ th subenvironment (or for the  $r$ th subskill) it is possible to specify a strategy.

Now Figure 17 represents a feasible system if there is enough information about  $L_0$  to determine a 'structured skill' (in particular to reduce the skill to sets of problems  $x^* \in X_r^* \subset X_r$  concerned with the  $r$ th subskill wherein if, for each  $r$ , there is a 'simplifying' procedure that renders any  $x^* \in X_r^*$  a more intelligible simplified problem, denoted by a stimulus  $x \in X_r - X_r^*$ ). If so, it can be shown that a pair of strategies  $C_1$  and  $C_2$  will maximize the rate of directed self-organization in the system  $J^*$  of the model of 2.6, namely:

$C_1$  given that  $X_p$  is selected, hence  $J_p$ , vary the simplification of an arbitrary sequence of  $x$  selected from  $X_p$  to minimize the

degree of simplification and to maximize  $\Delta\theta_p$ . If this proves impossible in  $J_p$  inform  $\Lambda$  and institute  $C_2$ .

$C_2$  Let  $\Lambda$  select  $X_r$ , hence  $J_r$ , such that the expected value of  $\Delta\theta$  is maximized. Proceed with  $C_1$ .

There is some evidence to indicate that  $C_1$  and  $C_2$  are, in a certain sense, 'optimal' strategies that maximize learning rate, but for most skills a much more elaborate hierarchical adaptive mechanism is needed to realize  $C_1$  and  $C_2$ . Note that although the selections made by strategy  $C_1$  are expressions in  $L_0$ , the selections made by  $\Lambda$ , embodying strategy  $C_2$ , are expressions in a machine metalanguage,  $L_1$  say, since they instruct the student to direct his attention to a different subskill<sup>75</sup>.

In Figure 18 a similar adaptive device is used to minimize the error rate in an inspection skill. The adaptive machine in Figure 18 injects 'false' signals into the inspector's display, senses whether he misses the signal, and adjusts the injected signal distribution to minimize this error rate. All 'false' signals are automatically excluded from the system output.

These machines have the status of catalysts that control an evolutionary process in the student's brain. They could be extended (by providing a capability for building an hierarchy  $L_1, L_2, \dots$  of metalanguages) so that this process became increasingly exteriorized (the machine could act as a medium, comparable to the brain, in so far as it manipulated 'concepts' like the student, but due to co-operation there is a possibility of a man machine compromise about the form of 'concept').

If the man is removed from this extended system, it becomes an undiluted 'artificial intelligence'. The 'induction principles' demanded by Church<sup>76</sup>, Markoff<sup>77</sup>, and others in an intelligent system must stem from the evolutionary rules that constrain the machine. Given that these are specified rightly, the 'artificial intelligence' will solve problems as man solves them. One method of embedding these rules might be to allow a learning machine to mature in contact with a group of men and a common linguistic environment. The other possibility is to apply constraints upon the machine which may be

(a) Heuristics determining classes of solution or classes of algorithms.

(b) Cost functions, determining how much the machine has to 'pay' in terms of a necessary and limited commodity, to maintain different structures or to apply different algorithms.

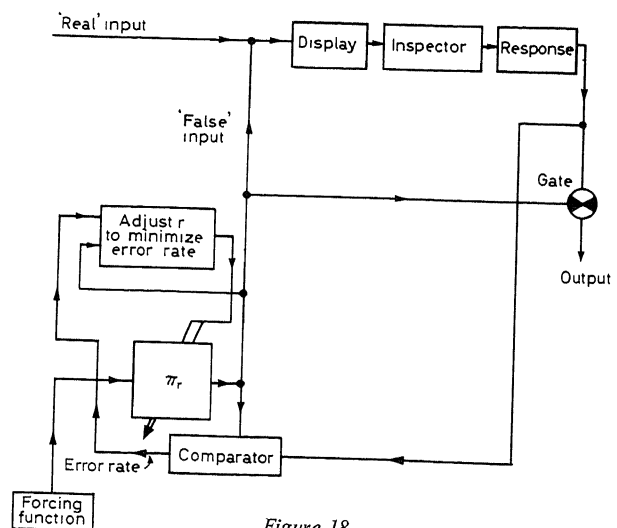


Figure 18



Some features of the better-known intelligent learning machines are now briefly outlined:

(1) One of the less elaborate systems is Manfred Kochen's<sup>78</sup> problem solver which applies various search strategies in a situation proposed by Bruner, Goodnow and Austin<sup>79</sup> in their studies of problem solving and learning. Its behaviour compares reasonably with human behaviour in the restricted experimental environment of these studies but the system is not intended to do more than this. This learning machine is an adaptive device that modifies the parameter,  $\phi$ , of the function set defined by its strategies.

(2) The hierarchy of control (in a symbolic identification of languages) is isomorphic with the structure of a list programme. Hence problem solving machines like Newall, Shaw, and Simon's G.P.S.<sup>80-82</sup> are abstractive learning machines. Variation of the hierarchical structure takes place as a function of the experience gained by applying the listed tests. Learning has at least a couple of different forms. (a) Learning to apply a test which entails the development of a local metric over the choice tree (equivalent to adaptation at a given level as in Fiegenbaum and Samuel's device<sup>118</sup>). (b) Learning to order tests and modify lists, which amounts to changing the hierarchical structure.

Minsky<sup>83, 84</sup> has particularly emphasized the role of heuristics or broad principles of problem solving that are built into a machine and guide it in selecting the algorithm it applies to its working data and to its own structure. Unless these are built into the device there seems no hope that any sort of order will appear. The heuristic constitutes an assertion of how a species of learning machine will look at its environment.

(3) Together with Minsky, Marzocco, Travis<sup>85</sup>, and others at S.D.C. take the very reasonable view that the most important heuristic determines a principle of restricted generalization whereby a given machine learns to apply a 'generalized' form of a previously successful algorithm when it is subsequently presented with a 'similar' problem. The crux of the learning process is to assign a connotation to the words in inverted commas.

This is also true of Amarel's<sup>86</sup> 'theorem proving' machine that 'learns to learn' in a limited domain (it proves 'theorems' that are mappings from a finite product set of elements into one coordinate of this set). Even here the search process (which entails developing choice trees and building up lists of descriptions of these procedures) is impracticable unless it is somehow restricted. Amarel urges (a) the need for a variety of descriptive metalanguages  $L_0, L_1, \dots$  in any intelligent learning machine (b) the need for some kind of economic control over the search process. Apart from obvious measures like informational value that can be used to characterize the result of a test, the machine must have to pay for making tests and maintaining the test systems in terms of a restricted commodity like the food or money discourse in 2.7. A similar point is made by Solomonoff<sup>87</sup>.

(4) Maron<sup>23</sup>, points out that the basic problem of artificial intelligence is choice. The logical rules that are built into a machine and the probabilistic bias it may acquire by learning, determines only what it may not do (the set of alternatives it is allowed to contemplate and those it may not contemplate). Nothing is said about what it should do within the contemplated set (an instruction to 'throw a dice' indicates our indifference but the act is realized by a specific and often inept 'decision procedure'). A machine that genuinely learns to solve problems must learn to choose and is likely to develop preferences and oddities.

### 3. The Mechanism in which Learning Machines are Realized

#### 3.1. Brain Analogies

(1) The brain is probably the best immediately available guide to how the various constraints should be realized by the control engineer. This realization is the concern of Bionics or Biological Cybernetics. Many of the hoary contentions of this field, like how 'random' or how 'specific' a brain may be, are losing their edge. Nowadays, few people subscribe to the doctrine of 'random' networks and few people imagine that the brain is a large 'telephone' system. There remain a number of less extremist biologically oriented design principles for learning machines that appear to have more than polemic value. These must be applied with the recognition that other than biological brains may have entirely different optimal designs.

(2) Brains do (and one suspects that learning machines must) have a well defined overall plan. The  $L_0$  informational interface\* between the organism and its environment is characterized, at the input, by an hierarchically arranged filter that extracts perceptual attributes. At the output, there are hierarchically controlled motor repertoires that give rise to sequences of coherent actions such as 'speech' or 'eating' each of which can be regarded as a whole. Between these specialized regions, there are other more or less organized entities concerned with integration and various forms of control. Probably the most comprehensive cybernetic model is provided by Anohkin<sup>88</sup>.

Within such a broad plan the brain has well defined functional units such as the various distinct memory systems, discussed, for example, by Brown<sup>89</sup> and Barbizet<sup>90</sup>.

(3) Very specific filtering networks have been constructed by von Foerster<sup>91</sup> and by Harmon<sup>92</sup>. This work has been stimulated and guided by the physiological investigations of Lettvin, Maturana, McCulloch and Pitts<sup>93</sup>, of Hubel and Wiesel<sup>94</sup> of Reichart<sup>95</sup> and others. Novikoff<sup>96</sup> has outlined principles for designing attribute filters on the basis of integral geometry. An axiomatic specification of a pattern is given by Lerner and Vashnik<sup>123</sup>.

At a more macroscopic level, Lerner<sup>121</sup> has proposed reflex analysing mechanisms which suggest and have been embodied in machine organizations. Finally, it can be argued that there is an information theoretical analogy between the filters of a learning machine and certain forms of communication and computation channels. Agalides<sup>98</sup> considers the matter at a microscopic level and Broadbent's<sup>99</sup> model (which is interpretable as a psychological hypothesis) refers to the macrostructure of a learning machine.

(4) Some of the filters in a brain are adaptive and some are predictive. On the basis of this analogy, Barlow and Donaldson<sup>100</sup> have built a highly specific coding system which adapts to minimize the redundancy in an input sequence which could be plausibly derived from an attribute filter. Uttley's<sup>101</sup> two variable conditional probability theory is interpretable in terms of a neurological mechanism through an identification between eqns (9) and (10) of 2.3 and some physiological process.

\* NOTES. The  $L_1$  or  $L_2$  description interface is normally different. A man or any learning machine has a definite location with reference to a given descriptive language. If this language is not specified he or it has an undetermined boundary. Man is bounded in one way as a social being and in another way as a food eating creature.

He is, at the moment, testing this hypothesis and the view that dendritic fields in the brain provide the connectivity of a predictive network. Almost all of the adaptive filters have some claim to biological precedent. Thus Rosenblatt's Perceptron was inspired by Hebb's learning theory and assemblies of the kind that Taylor and Steinbuch have built are consonant either with Hebb's view<sup>1, 2</sup> or more specific theories such as those of Milner<sup>103</sup>.

In brains, the hierarchical structure of control is partly apparent in the anatomy. Bishop<sup>104</sup> points out that the Broad Pattern (which is anatomically defined) has evolutionary origins.

Braynes, Napalkov and Sechvinsky<sup>105</sup>, and Napalkov<sup>97</sup>, have developed a particularly elegant description of conditioned reflexes in terms of algorithms. In many animals it is possible to exhibit the hierarchy of control as an interaction between different levels of algorithms which can, of course, be rigorously specified as a relation between normal algorithms in the sense of Markoff. The hierarchy of algorithms describes a structure of reflexive chains which have, at any moment, a definite location, but there is no reason why the description should always refer to the same physical structure. Hence the hierarchy of algorithms is an organization which sometimes may and sometimes may not reflect an anatomical hierarchy. These authors have built several machines to illustrate their own ideas and the obviously related concepts of the response hierarchy biologists such as Tinbergen<sup>107</sup>. One of the most comprehensive hierarchical learning systems is described by Andrea<sup>41</sup>.

(5) The brain is a parallel computing mechanism and it has a great deal of structural and functional redundancy. Work such as Cowan's<sup>138</sup> on error correction in computing networks leads to an image (the design of a typical error correcting, stable, network) which is also highly redundant and which closely resembles the connectivity of a brain. In this kind of system, the same function may be mediated by many different components in many different locations. The brain is many physical realizations of a computing system rather than a single realization (a brain is functionally similar to a computing system—it bears little resemblance to the object called a computer). The minimal components are not always well defined and the attributes of physical events that count as signals are rarely well defined (thus in some parts of a brain, neurone groups act as the minimal subsystems, in other parts, single neurones, and in others the components of a neurone. Similarly, 'signals' may be impulses or the phase relations between impulses or the mean rate of impulses, or analogue variables that are not directly related to impulses). It is not unreasonable to specify this kind of microscopic redundancy of function in the components of any learning machine. Certain obvious advantages are obtained. The trouble is that the idea of a 'component' and the familiar distinction between signals and objects that modify these signals is lost.

Finally, there is a redundancy at the level of macroscopic properties. It is probably pointless to ask what in a brain mediates its memory. An indefinitely large number of physical mechanisms might (and in the sense of being history dependent) must act in this capacity. The real issue is to discover which are relevant to the computations that interest us. Thus a brain can have memory in terms of excited loops of neurones, in terms of a distribution of facilitation that favours the propagation of a particular wave of activity, in terms of specific macromolecular changes at the synaptic interface and possibly in terms of local and specific changes in messenger R.N.A. There is no reason to

suppose that any one of these is responsible for any facet of memory and their relative importance probably depends upon the state of the brain. Given the component specification suggested a moment ago, similar properties may be expected in a learning machine.

(6) The brain includes an attention mechanism, chiefly embodied in the reticular formation which, according to McCulloch and Kilmer<sup>109</sup> can be regarded as an input and output transforming system with parameters controlled by the higher centres. It determines the level in an hierarchy of control at which a feedback loop, involving the data occupying the brain's attention, will be completed. Ideally this is the level of abstraction suited to a given type of data. It has already been argued, on entirely different grounds, that an attention mechanism is needed in any competent learning machine.

(7) The fact that a brain is a collection of closely coupled evolutionary fragments, does not appear to be accidental, nor is the fact that the ontogeny of a brain recapitulates the phylogeny of the brain of a species as the system undergoes the process of maturation. Such manifestations as imprinting (that certain behaviours are acquired by the organism if and only if a specific state of its environment occurs at a specific point in its development) are a predictable consequence of maturation and unless an alternative is suggested, any realistic learning machine must also undergo a process of maturation as proposed in 2.8 in its normal environment.

In this connection, Young's investigation<sup>122</sup> of the octopus has recently revealed that the brain of this animal is an hierarchically organized homeostatic system in which the hierarchical levels are related to the sensory and motor systems especially characteristic of distinct evolutionary stages (touch, visual distance receptors, and visual pattern systems). The octopus is peculiar in that the functions of its brain are localized (and, hence, are readily discriminated) whereas in most animals the functions are distributed in a way which obscures the overall plan. However, one may hypothesize that the organization exposed by Young's work persists, even when its embodiment is hard to discern as it is in a mammalian brain.

(8) There are regions in a brain that could, very readily, act as the medium in which dynamic organizations evolve. Relatively undifferentiated parts of the cerebral cortex display an activity like Beurle's networks and although it is possible that evolutionary processes, in the sense given in 2.8, occur in almost any biological system, it is tempting to suppose that these parts are specially developed as a 'medium'. There is plenty of evidence that evolutionary processes, in the sense given in 2.8, go on in brains, that they underlie the learning behaviour of the associated organism and serve to extrapolate its biological evolution into the field of language and social interaction. It has been argued, in less flowery terms, that much the same comments may apply to learning machines.

### 3.2. *Non-biological Embodiment*

Since useful learning machines are certainly large, their components (with the reservation of (5) above upon using this word) must be small. There seems no reason why one or another of the microminiaturization techniques (such as the techniques discussed by Shoulders) cannot be used to fabricate the perceptual filters and motor hierarchies of (2), (3) and (4) above in conventional circuitry. The fabrication need not be perfect, since the system is stable against crass structural disturbances. These



methods are not able to realize a competent version of maturation [required by (7) above] or of a medium [as in (8)]. So far as maturation is concerned, the constructing device for etching or embedding the circuit would have to be controlled as a function of the interaction between the brain it is producing and the normal environment of this brain and although this is not impossible, the expedient seems, at first sight, to be unduly cumbersome.

Various attempts have been made, chiefly by MacKay and in my own laboratory, to 'grow' connectivity in a network using metallic threads or dendrites that develop by electrodeposition in a form determined by the activity induced through the connections they instantaneously determine. Whereas MacKay<sup>110</sup> has brought to competence a method of 'growing' dendrites that act as connective elements, our own work yielded a very crude but somewhat wider application of this process. Briefly, it was possible to construct, from dendrites and capacitors alone, a realized network of Crane's<sup>111, 112</sup> 'neuristors' (a 'neuristor' is an active transmission line, of which a nerve fibre is a special case. Crane has shown that suitably coupled networks of neuristors are able to compute any Boolean or Probabilistic function. Their realization, using Lilley's passivated wires, is considered by Stewart<sup>113</sup>). In the present system, instable dendrites subverted the non-linear switching action of a neuristor whereas stable connective dendrites acted as the coupling and charging impedances. If the system is chosen so that the deposition is partly reversible, the dendrites have a reproductive characteristic and competitive and co-operative interaction takes place between members of a population of developing dendrites<sup>4, 26, 59</sup>. Since, as Feldman<sup>114</sup> points out, a neuristor system is ideally suited to act as a medium for evolving dynamic organizations, the dendrite realization of a neuristor network appears particularly attractive and it may be possible to use this system now that more elegant procedures have been developed.

The basic criticism is that the size is too large. Ideally, the active transmission lines should be large macromolecules and Bowman<sup>115</sup> argues that certain polymers could be used for this purpose. To complete the specification, it is necessary to transmit signals along polymer 'neuristors' whilst the polymerization is in progress, and to use these signals through a local catalytic action to control the further extension of the 'neuristor' network. Physical chemists seem willing to countenance the possibility of controlled catalysis, at least in principle.

## References

- <sup>1</sup> GABOR, D., WILBY, W. P. L. and WOODCOCK, R. A self-optimizing non-linear filter predictor and simulator. *Proc. 4th London Symp. on Inf. Theory*
- <sup>2</sup> McCULLOCH, W. S. and PITTS, W. A logical calculus of the ideas immanent in nervous activity. *Bull. math. Biophys.* 5 (1943) 115
- <sup>3</sup> MACKEY, D. M. Comments on Terminology in *Aspects of Artificial Intelligence* (Ed. C. Muses). 1962. Plenum Press
- <sup>4</sup> PASK, G. Physical analogues to the growth of a concept. *The Mechanisation of Thought Processes*. 1958. London; H.M.S.O.
- <sup>5</sup> DEUTSCH, J. A. *The Structural Basis of Behaviour*. 1957. Cambridge University Press
- <sup>6</sup> ANGYAN, A. J. 'An Analogue Model to Demonstrate Some Aspects of Neural Adaptation.' *The Mechanisation of Thought Processes*. 1958. London; H.M.S.O.
- <sup>7</sup> WALTER, W. GREY *The Living Brain*. 1953. London; Duckworth
- <sup>8</sup> HARMON, L. D. The artificial neurone. *Science*, 129 (1959) 962
- <sup>9</sup> BEURLE, R. L. *Activity in a block of cells capable of regenerating impulses*, R.R.E. Memoranda, 1042 and 1043, 1954
- <sup>10</sup> BEURLE, R. L. Storage and manipulation of information in the brain. *J. Inst. elect. Engrs*, 5, New Series (1959)
- <sup>11</sup> UTTLEY, A. M. *The conditional probability of signals in the nervous system*. R.R.E. Memorandum No. 1109, 1955
- <sup>12</sup> UTTLEY, A. M. A theory of the mechanism of learning based on the computation of conditional probabilities, *Proc. 1st Congr. Int. Assoc. Cybernetics*, Namur, 1956. Paris; Gauthier Villars
- <sup>13</sup> UTTLEY, A. M. Conditional Probability Computing in the Nervous System. *The Mechanisation of Thought Processes*. 1958. London; H.M.S.O.
- <sup>14</sup> UTTLEY, A. M. The Engineering Approach to the Problem of Neural Organisation, *Progress in Biophysics*, Vol. II. 1961. Oxford; Pergamon Press
- <sup>15</sup> UTTLEY, A. M. Conditional probability machines and conditional reflexes, *Automata Studies* (Eds. Shannon & McCarthy). 1956. Princeton Univ. Press
- <sup>16</sup> ASHBY, W. ROSS *Design for a Brain*. 1959. London; Chapman & Hall
- <sup>17</sup> AIZERMAN, M. A. Automatic control learning systems, *2nd I.F.A.C. Conference*, Basle 1963
- <sup>18</sup> PASK, G. The cybernetics of evolutionary and self-organising systems. *Conference Generale at 3rd Congr. Int. Assoc. for Cybernetics*, Namur 1961. Paris; to be published by Gauthier Villars
- <sup>19</sup> GREENE, P. H. Computers that perceive, learn and reason, *General Systems Yearbook*, Vol. 4
- <sup>20</sup> ANDREW, A. M. Learning Machines, *The Mechanisation of Thought Processes*. 1958. London; H.M.S.O.
- <sup>21</sup> MINSKY, M. and SELFRIDGE, O. Random nets, *Proc. 4th London Symp. on Information Theory* (Ed. C. Cherry). 1962. London; Butterworth
- <sup>22</sup> ASHBY, W. ROSS *Design for an intelligence amplifier. Automata Studies* (Eds. Shannon & McCarthy). 1954. Princeton Univ. Press
- <sup>23</sup> MARON, H. E. *Artificial intelligence and brain mechanisms*, Memorandum RM 3522 PR, Rand Corporation, 1963
- <sup>24</sup> MARON, H. E. The design principles for an intelligent machine. *Int. Symp. on Inf., Inst. Radio Engrs Trans. in Inf. Theory*, Vol. II, 8, No. 5 (1962)
- <sup>25</sup> STEINBUCH, R. Learning matrices, *Kybernetik* (1961) 36
- <sup>26</sup> PASK, G. *An Approach to Cybernetics*. 1961; Hutchinsons Statistical computation and statistical automata. *Proc. DAGK Conf.* Karlsruhe, 1963
- <sup>27</sup> TAYLOR, W. K. Theory of cortical organisation and learning, *Int. Symp. Inf. Theory. Inst. Radio Engrs Trans.*, Vol. II, 8 (1962); A pattern recognizing adaptive controller. Automatic and Remote Control. 1963. London; Butterworths. Munich; Oldenbourg
- <sup>28</sup> WIDROW, B. Information Storage in a Network of Adaline Neurones, *Self-organising Systems*. 1962. Spartan Press
- <sup>29</sup> WILLIS, G. D. *Plastic neurones as memory elements*, Lockheed Report, LMSD 48432, 1959
- <sup>30</sup> BABCOCK, M. *An adaptive reorganising automaton*, Tech. Rep., NONR 1834 (21) E.E.R.L., Univ. Ill. 1961
- <sup>31</sup> ROSENBLATT, F. *Principles of neurodynamics, perceptrons, and the theory of brain mechanisms*. Cornell Aeronautical Lab. Rep. No. VG-1196-G-8, 1961
- <sup>32</sup> PAPERT, S. *Redundancy and linear logical nets*. Bionics Symposium, WADD Tech. Rep., 60-600, Sept. 1960
- <sup>33</sup> SINGLETON, S. A Test for Linear Separability as Applied to Self-organising Machines, *Self-organising Systems* (Eds. Yovits, Jacobi & Goldstein). 1962. Spartan Press
- <sup>34</sup> CAMERON, S. *An estimate of the complexity requisite in a universal decision network*, WADD Bionics Symposium, Washington, ASTIA

- <sup>35</sup> IVANHENKO, A. G. *Avtomatika*, Kiev, 1963
- <sup>36</sup> WILLIS, G. D. The functional domain of complex systems, *Principles of Self Organisation* (Eds. Zopf & von Foerster). 1962. Pergamon Press
- <sup>37</sup> ASHBY, W. ROSS *Introduction to Cybernetics*. 1957. London; Chapman & Hall
- <sup>38</sup> HAIRE, P. F. and HAROULESS, G. *Jenny, an imperfect homeostat*, AFCRC TN 60379, April 1960; WILLIAMS, R. E. *Static and dynamic responses of the homeostat Jenny*, AFCRL 505, June 1961
- <sup>39</sup> MACKAY, D. M. Mindlike behaviour in artefacts, *Brit. J. Phil. Sci.* 2 (1951) 105-21
- <sup>40</sup> MACKAY, D. M. The Epistemological Problem for Automata. *Automata Studies* (Eds. Shannon & McCarthy). 1956. Princeton Univ. Press
- <sup>41</sup> ANDREA, J. H. Stella, a scheme for a learning machine. *2nd Congr. I.F.A.C.*, Basle 1963
- <sup>42</sup> PASK, G. Comments on an indeterminacy that characterises a self organising system. *Proc. Spring School of Theoretical Physics Cybernetics of Neural Processes*, Naples, 1962
- <sup>43</sup> PASK, G. Physical and linguistic evolution in self-organising systems. *I.F.A.C. Symp. on Self-Adaptive Systems*, Rome, 1962
- <sup>44</sup> FOERSTER, H. VON On Self-organising Systems and Their Environment. *Self-organising Systems* (Eds. Yovits & Cameron). 1960. London; Pergamon Press
- <sup>45</sup> FOERSTER, H. VON *Preorganisation for a self-organising system*. Univ. Illinois Rep. on Contract NONR 1834 (21), 1961
- <sup>46</sup> FOERSTER, H. VON and PASK, G. A predictive model for a self-organising system. *Cybernetica* 4 (1960); 1 (1961)
- <sup>47</sup> WATTANABE, S. Learning process and the inverse H theorem. *Int. Symp. on Inf. Theory, Inst. Radio Engrs Trans. Inf. Theory*, Vol. II, 8, No. 5 (1962)
- <sup>48</sup> BURKE, A. W. Computation, Behaviour and Structure. *Principles of Self-organisation* (Eds. Yovits & Cameron). 1960. Pergamon
- <sup>49</sup> NEUMANN, J. VON Unpublished works
- <sup>50</sup> LOEFGREN, L. *Qualitative limits for automatic self repair*, Tech. Note Contract NONR 1834 (21), E.E.R.L. Univ. of Illinois, 1961
- <sup>51</sup> LOEFGREN, L. *Tessellation models of self-repair*, Tech. Note Contract NONR 1834 (21), E.E.R.L., Univ. of Illinois, 1961; *Biological Prototypes and Synthetic Systems* (Eds. Bernard & Kare). 1962. Plenum Press
- <sup>52</sup> RASHEVSKY, N. *Mathematical Biophysics*. 1960. New York; Dover
- <sup>53</sup> ROSEN, R. The representation of biological systems from the standpoint of the theory of categories, *Bull. Math. Biophysics*, 20 (1958) 317-341
- <sup>54</sup> ROSEN, R. A logical paradox implicit in the notion of a self reproducing automaton, *Bull. Math. Biophys.* 21 (1959)
- <sup>55</sup> PASK, G. A Proposed Evolutionary Model, *Principles of Self-organisation* (Eds. G. Zopf & H. von Foerster). 1962. Pergamon Press
- <sup>56</sup> TODA, M. The Design of a Fungus Eater, *Behavioural Sciences*, 7 (1962) 164-183
- <sup>57</sup> FOERSTER, H. VON Miscellaneous reports under NONR Contract 1834 (21), E.E.R.L. Univ. of Illinois
- <sup>58</sup> CAIANIELLO, E. R. Outline of a theory of thought processes and thinking machines, *J. Theoret. Biol.*, 2 (1961) 204-235
- <sup>59</sup> PASK, G. The growth process in a cybernetic machine. *Proc. 2nd Conf. Int. Assoc. Cybernetics*, Namur, 1958
- <sup>60</sup> FOERSTER, H. VON *Biologic, Biological Prototypes and Synthetic Systems* (Eds. Bernard & Kare). 1962. Plenum Press
- <sup>61</sup> SELFRIDGE, O. *Pandemonium*, *Mech. of Thought Processes*. 1958. London; H.M.S.O.
- <sup>62</sup> PRINGLE, J. W. S. On the parallel between learning and evolution. *Behaviour*, 3 (1951) 174
- <sup>63</sup> FARLEY, B. and CLARKE, R. Activity in a network of neurone-like elements. *Proc. 4th London Symp. on Inf. Theory* (Ed. C. Cherry). 1962. Butterworth
- <sup>64</sup> WIENER, N. *Cybernetics*, 2nd ed. 1962. Wiley
- <sup>65</sup> WIENER, N. *Comments at Spring School of Theoretical Physics*, Naples 1962 (to be published)
- <sup>66</sup> TUSTIN, A. *The Mechanism of Economic Systems*. 1953. Harvard Univ. Press
- <sup>67</sup> CHERRY, C. *On Human Communication*. 1956. New York; Wiley
- <sup>68</sup> GEORGE, F. H. *The Brain as a Computer*. 1961. London; Pergamon Press
- <sup>69</sup> GEORGE, F. H. Pragmatic machines, *Proc. DAGK Conf. on Cybernetics*, 1963. Karlsruhe
- <sup>70</sup> EDWARDS, W. *Probabilistic information processing in command and control systems*. AD 3789-12-T. Engng Psychol. Lab. Univ. Michigan. March, 1963
- <sup>71</sup> PASK, G. Teaching machines, *Proc. 2nd Congr. Int. Assoc. Cybernetics*, Namur, 1958
- <sup>72</sup> PASK, G. The teaching machine as a control mechanism, *Trans. Soc. Inst. Tech.* June 1960
- <sup>73</sup> PASK, G. The logic and behaviour of self-organising systems as illustrated by the interaction between men and adaptive machines, *Int. Symp. on Inf. Theory*, Brussels, 1962
- <sup>74</sup> PASK, G. Self-organising systems involved in human learning and performance. *Bionics Symp.*, Dayton, Ohio, 1963
- <sup>75</sup> PASK, G. *A model of learning applicable within systems stabilised by an adaptive teaching machine*. Tech. Note 1, USAF Cont. AF 61 (052)-402, 1962
- <sup>76</sup> CHURCH, A. *Introduction to Mathematical Logic*. 1944. Princeton Univ. Press
- <sup>77</sup> MARKOFF, A. A. *Theory of Algorithms*. 1954. Moscow; Acad. Sciences
- <sup>78</sup> KOCHEN, M. Experimental study of 'Hypothesis formation' by computer, *Proc. London Symp. Inf. Theory* (1960). 1962. Butterworths
- <sup>79</sup> BRUNER, J. S., GOODNOW, J. J. and AUSTIN, G. A. *A Study of Thinking*. 1956. New York; Wiley
- <sup>80</sup> NEWALL, A., SHAW, J. C. and SIMON, H. A. Elements of a theory of human problem-solving. *Psychol. Rev.*, 65 (1958) 152-166
- <sup>81</sup> NEWALL, A., SHAW, J. C. and SIMON, H. A. Report on a general problem-solving programme. *Proc. Int. Conf. Inf. Processing*, Paris, UNESCO, 256-264
- <sup>82</sup> NEWALL, A., SHAW, J. C. and SIMON, H. A. A variety of intelligent learning in a general problem solver. *Rand Corporation, Tech. Rep.* 1959
- <sup>83</sup> MINSKY, M. Steps towards artificial intelligence, *Proc. Inst. Radio Engrs*, 49 (1961) 8-30
- <sup>84</sup> MINSKY, M. Some Methods of Artificial Intelligence and Heuristic Programming, *Mech. of Thought Processes*, N.P.L., 1958. H.M.S.O.
- <sup>85</sup> MARZOCCO, F., TRAVIS, L. et al. SDC Documents 1962
- <sup>86</sup> AMAREL, S. *An Approach to Automatic Theory Formation Self-organising Systems*, 1962 (Ed. Yovits, Cameron, Jacobi). 1962. Spartan Press
- <sup>87</sup> SOLOMONOFF, R. J. *Research in inductive inference for the Period April 1, 1959 to Nov. 30, 1960*. Progr. Rep. ZTB 139, Con. AF (638)-376, 1961
- <sup>88</sup> ANOHKIN, P. *Proc. of Symp. on Inf. Processing of Nervous System*. Leyden, 1962. To be published
- <sup>89</sup> BROWN, J. Information, Redundancy and the Decay of the Memory Trace, *Mech. of Thought Process* N.P.L., 1958. H.M.S.O.
- <sup>90</sup> BARBIZET, J. and ALBARDE, P. Memoires humaines et memoires artificiels, *Concours med.*, 6 (1961)
- <sup>91</sup> FOERSTER, H. VON *Circuitry of Clues to Platonic Ideation Aspects of Artificial Intelligence* (Ed. C. Muses). 1962. Plenum Press

- <sup>92</sup> HARMON, L. D. Studies with Artificial Neurones. I. Properties and Functions of an Artificial Neurone. *Kybernetik* 1, 3 (1962) 89-101
- <sup>93</sup> LETTVIN, MATTURANA, MCCULLOCH and PITTS What the frog's eye tells the frog's brain. *Proc. Inst. Radio Engrs*, Nov. (1959)
- <sup>94</sup> HUBEL, D. H. and WIESEL, T. N. Receptive fields of single neurones in the cats striate cortex, *J. Physiol.* 148 (1959) 574-591
- <sup>95</sup> REICHART, W. *Symposium on Information Processing in Nervous System*, Leyden. 1962
- <sup>96</sup> NOVIKOFF, A. Integral Geometry in Pattern Perception, *Principles of Self-organisation* (Ed. Foerster and Zopf). 1962. Pergamon
- <sup>97</sup> NAPALKOV, A. A study of the laws of development of complex conditioned reflex systems. *Vest. Nik. Moscow Univ.* 2 (1958)
- <sup>98</sup> AGALIDES, E. *Communication and information theory aspects of the nervous system*, Tech. Status Report General Dynamics Corp., April 1963
- <sup>99</sup> BROADBENT, D. S. *Perception and Communication*. 1957. Pergamon
- <sup>100</sup> BARLOW, B. and DONALDSON, P. Sensory Mechanisms, the Reduction of Redundancy and Intelligence, *The Mech. of Thought Processes*, N.P.L. 1958. H.M.S.O.
- <sup>101</sup> UTTLEY, A. M. Properties of plastic networks. *Biophys. J.* 2 (1962) 2
- <sup>102</sup> HEBB, D. O. *The Organisation of Behaviour*. 1949. New York; Wiley
- <sup>103</sup> MILNER, P. M. The cell assembly, Mk. II. *Psychol. Rev.* 64 (1957) 242-252
- <sup>104</sup> BISHOP, G. Environmental Feedback in Brain Function, *Self-organising Systems* (Eds. Yovits & Cameron). 1960. Pergamon
- <sup>105</sup> BRAYNES, NAPALKOV and SECHVINSKY *Problems of cybernetics*, D.S.I.R. translation 1961
- <sup>106</sup> BRAYNES and SECHVINSKY *Matrix structures and simulation of learning*, *Trans. Inst. Radio Engrs Symp. on Inf. Theory*, Brussels, 1962
- <sup>107</sup> TINBERGEN, N. *A Study of Instinct*. 1951. Oxford Univ. Press
- <sup>108</sup> COWAN, J. Many-valued Logics and Reliable Automata, *Principles of Self-organisation* (Ed. von FOERSTER & Zopf). 1962. Pergamon
- <sup>109</sup> MCCULLOCH and KILMER Address given to Bionics Symposium at Dayton, Ohio, 1963
- <sup>110</sup> MACKAY, D. M. and AINSWORTH, A. Electrolytic growth processes. *DAGK Conf.*, Karlsruhe, 1963.
- <sup>111</sup> CRANE, H. D. *Neuristor studies*, Tech. Rep. 1506, Stanford Electronics Laboratories, 1960
- <sup>112</sup> CRANE, H. D. The Neuristor, *Principles of Self-organising Systems* (Eds. von Foerster & Zopf). 1962. Pergamon
- <sup>113</sup> STEWART, R. M. *Theory of Structurally Homogeneous Logic Nets Biological Prototypes and Synthetic Systems*, Vol. 1, Bernard & Kare. 1962. Plenum Press; Personal communication and talks at 2nd Conf. Int. Assoc. Med. Cybernetics at Amsterdam, 1962
- <sup>114</sup> FELDMANN, R. A homogeneous network approach to self-structuring systems. *Microwave Res. Inst.*, Brooklyn Polytechnic, N.Y. 1961
- <sup>115</sup> BOWMAN, R. A New Transmission Line Leading to a Self-Stabilising System, *Principles of Self-Organisation* (Eds. von Foerster & Zopf). 1962. Pergamon
- <sup>116</sup> RATZ, R. C. and THOMAS, G. M. *The Development of a Conditional Probability Computer for Control Application. Information Processing* (Ed. C. M. Popplewell). 1963. North Holland Publishing Company
- <sup>117</sup> MESAROVIC, M. and BANERJI, B. A computer programme for the generation of new concepts from old. *Proc. DAGK Symposium*, Karlsruhe, 1963
- <sup>118</sup> FIEGENBAUM, SIMON, H. *Elementary Perceiving and Memorising Machine. Information Processing* (Ed. C. M. Popplewell). 1963. North Holland Publishing Company
- <sup>119</sup> CHICHINADZE, C. A Contribution to the Use of Self-adjusting Systems for the Mechanical Synthesis of Control Systems. *Automatic and Remote Control*. 1963. London; Butterworths. Munich; Oldenbourg
- <sup>120</sup> LERNER, A. YA. and VASHNIK, V. N. Recognition of patterns with the aid of generalized portraits. *Automat. and Telemekh.* XXIV, 6 (1963)
- <sup>121</sup> LERNER, A. YA. Miscellaneous reports on perceptual filters (in Russian) Acad. Sci., Moscow, 1960-62
- <sup>122</sup> YOUNG, J. Z. Some essentials of neural memory systems. *Proc. 10th Int. Congr. Electronics*. Rome, 1963
- <sup>123</sup> ULLMAN, L. R. Cybernetic models that learn sensory motor connections. *Medical Electronics and Biological Engineering* 1, No. 1 (1963)
- <sup>124</sup> TAYLOR, W. K. A pattern recognizing adaptive controller. *Automatic and Remote Control*. 1963. London; Butterworths. Munich; Oldenbourg

# Invariance of Sampled-data and Adaptive Sampled-data Systems

V. M. KUNTSEVICH and YU. V. KREMENTULO

## Summary

In the report it is shown that the invariance conditions can be used successfully in usual continuous control systems and also in sampled-data control systems. This is especially important in adaptive impulse systems. In the first section of the report fundamentals of invariance theory of combined sampled-data systems are considered. For such systems the invariance conditions of the controlled variable were obtained both for sampling intervals and for any period of time.

Possibility of a significant improvement of adaptive systems based on the invariance theory is investigated. The impulse extremal system of the incremental type with a test signal is considered. It is shown that by introducing a certain change in the structure of the adaptive system and according to the choice of parameters, a complete elimination of the system error due to a definite group of disturbances which displace the extremum points along the axis of controlled variable, is possible. In this case the errors due to these disturbances are eliminated and the region of stability is considerably expanded.

## Sommaire

Dans ce rapport on montre que les conditions d'invariance définies dans les systèmes continus peuvent être appliquées avec succès aux systèmes de commande échantillonnés. Ce point est particulièrement important dans les systèmes adaptatifs impulsifs. La première partie de ce rapport traite des bases fondamentales d'application de la théorie de l'invariance aux systèmes échantillonnés mixtes; les conditions d'invariance des variables commandées sont déterminées durant les intervalles d'échantillonnage et pour toute période de temps.

On examine ensuite la possibilité d'améliorer les systèmes adaptatifs sur la base de la théorie de l'invariance. On considère un système extrémalisant impulsif du type à incréments avec signal d'essai. On montre qu'en modifiant la structure du système adaptatif et en effectuant un certain choix de paramètres, on obtient une élimination complète de l'erreur due au déplacement de l'extrémum, ce déplacement de l'extrémum étant dû lui-même à une catégorie de perturbations bien définies. Le domaine de stabilité se trouve considérablement étendu.

## Zusammenfassung

Der Aufsatz zeigt, daß sich die Invarianzbedingungen (Kompensation von Störeinflüssen) sowohl in den üblichen kontinuierlichen Regelsystemen als auch in Abtastregelsystemen erfolgreich verwendet lassen. Dies ist besonders für selbst-einstellende Abtastsysteme (Impulssysteme) wichtig. Zunächst gilt die Betrachtung den Grundlagen der Invarianztheorie kombinierter Abtastsysteme. Die Invarianzbedingungen für die Regelgröße solcher Systeme werden sowohl für die Abtastperiode als auch für jeden anderen Zeitabschnitt abgeleitet.

Es wird untersucht, ob eine wesentliche Verbesserung der selbst-einstellenden Regelsysteme auf Grund der Invarianztheorie möglich ist. Ein Extremalwert-Abtastregelsystem mit schrittweiser Näherung durch ein Testsignal wird betrachtet. Es zeigt sich, daß durch eine bestimmte Strukturänderung des selbst-einstellenden Systems und auf Grund der Wahl der Parameter eine völlige Ausschaltung des Systemfehlers möglich ist, sofern der Fehler einer bestimmten Klasse von Störungen angehört, die die Extremstellen entlang der Achse der Regelgröße verschieben. In diesem Fall werden die auf derartigen Störungen beruhenden Fehler ausgeschaltet und der Stabilitätsbereich beträchtlich erweitert.

One of the important scientific trends in the theory of automatic control is the theory of the construction of systems on the basis of compensation of the influence of disturbances, or the theory of invariance of the controlled value.

As is known, however, the invariance theory was recently used extensively only for ordinary continuous control systems<sup>1-7</sup>. Attempts were made in a number of works<sup>8-15</sup> to extend the general principles of this theory to sampled-data control systems, but there has not yet been any full and systematized statement of the invariance theory for such systems. The above statement relates in a still greater degree to adaptive systems in general and sampled-data systems of this type in particular. Since adaptive systems are a special type of non-linear systems, then, as will be shown below, the introduction of compounding disturbance links makes it possible not only to improve the quality of systems when compensating the influence of disturbances, but also to extend the stability region of these systems.

The authors consider that the main aim of their paper is to demonstrate the fact that the sampled-data system analysis and synthesis methods expounded below can serve as the basis for the construction of control systems with considerably greater accuracy than existing systems.

Henceforward the following constraints and assumptions will be accepted: (a) synchronous sampled-data systems with amplitude modulation are considered; (b) the sampling period  $T$  is constant; (c) the pulse element is ideal; (d) the equations are written in deviations; and (e) initial conditions are zero.

Since sampled-data systems of fairly complex structure will be considered, the consideration will begin with the method of solving the equations of multiloop sampled-data systems.

## Sampled-data Systems

### Multiloop Sampled-data Systems Equations

A number of works<sup>16-23</sup> have been concerned with the compilation and solution of the equations of sampled-data systems. The solution of the equations of multiloop sampled-data systems is given in the most general and convenient form by Burshtein<sup>17</sup>. The method suggested below has features in common with Burshtein's method, but allows one to avoid a number of intermediate operations and to simplify the calculations.

In the most general form the equation for the  $k$ th coordinate of a multiloop sampled-data system can be written thus:

$$x_k(s) = \sum_{i=1}^n W_{ki}(s) x_i(s) + \sum_{i=1}^n \sum_{j=1}^{l_{ki}} b'_{kij} x_i^*(z) b_{kij}(s) + \sum_{i=1}^m R_{ki}(s) F_i(s) + \sum_{i=1}^m \sum_{j=1}^{P_{ki}} C'_{kij} F_i^*(z) C_{kij}(s) \quad (1)$$

where  $x_k$ ,  $x_i$  are the coordinates of the system,  $F_i$  the external disturbances,  $n$ ,  $m$  the number of selected coordinates and external disturbances respectively,  $l_{k i}$ ,  $P_{k i}$  the number of parallel links (pulse-continuous) between the coordinate  $x_k$  and  $x_i$  and the coordinate  $x_k$  and the external disturbance  $F_i$  respectively;  $W$ ,  $b$ ,  $b'$ ,  $c$ ,  $c'$  and  $R$  are the corresponding transfer functions, shown in *Figure 1*, which depicts part of a multiloop sampled-data system ( $k$ th node).

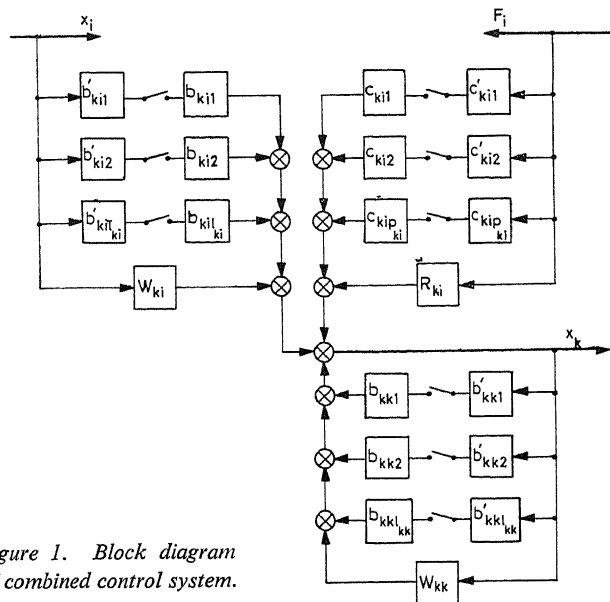


Figure 1. Block diagram of combined control system.

If one takes into consideration the additional coordinates:

$$\begin{array}{l}
 b'_{1111}(s)x_1(s) = x_{n+1}(s) \\
 \cdot \quad \cdot \quad \cdot \quad \cdot \quad \cdot \quad \cdot \\
 b'_{11 \dots l_{11}}(s)x_1(s) = x_{n+l_{11}}(s) \\
 b'_{1n1}(s)x_n(s) = x_{n+l_{11} + \dots + l_{1(n-1)+1}}(s) \\
 \cdot \quad \cdot \quad \cdot \quad \cdot \quad \cdot \quad \cdot \\
 b'_{1n \dots l_{1n}}(s)x_n(s) = x_{n+l_{11} + \dots + l_{1n}}(s), \text{ etc.}
 \end{array} \tag{2}$$

then the equations of the multiloop sampled-data system can be given in an ordered form:

[illegible]

where

$$N = n + \sum_{k=1}^n \sum_{i=1}^n l_{ki}$$

is the full amount of coordinates of the system (including the additional ones),

$$A_k(s) = - \left[ \sum_{i=1}^m R_{ki}(s) F_i(s) + \sum_{i=1}^m \sum_{j=1}^{P_{ki}} C'_{kij} F_i^*(z) C_{kij}(s) \right],$$

$$a_{ki}(s) = W_{ki}(s); \quad a_{kk}(s) = W_{kk}(s) - 1; \quad \text{etc.}$$

and are numbered in accordance with (2).

System (3) formally contains  $N$  equations with 2  $N$  unknowns,  $x_j(s)$  and  $x_j^*(z)$ . As in ref. 17, the terms containing  $z$  transforms of the coordinates will be transferred to the right-hand side. The resulting system will be solved relative to the arbitrary coordinate  $x_j(s)$ . This gives:

$$x_j(s) = \frac{\Delta x_j(s)}{\Delta(s)} \quad (4)$$

where

$$\Delta(s) = \begin{vmatrix} a_{11}(s), \dots, a_{1N}(s) \\ \vdots \\ a_{N1}(s), \dots, a_{NN}(s) \end{vmatrix} \quad (5)$$

is the common determinant of a purely continuous system.

Eqn (6)

 \*

Determinant (6) may be presented in the form:

Eqn (7)

 †

The first of the determinants entering into (7) will be denoted by  $\Delta_A^j(s)$ , and the remainder by  $\Delta_{x_k}^{j*}(s)$ . Bearing in mind the notation adopted:

$$x_j(s) = \frac{\Delta_A^j(s)}{\Delta(s)} - \sum_{k=1}^N x_k^*(z) \left( \frac{\Delta_{x_k^*}^j(s)}{\Delta(s)} \right) \quad (8)$$

\* Eqn (6)

$$\Delta_{x_j}(s) = \begin{vmatrix} a_{11}(s), \dots, a_{1j-1}(s); A_1(s) - \sum_{j=1}^N a'_{1j}(s)x_j^*(z); a_{1j+1}(s), \dots, a_{1N}(s) \\ \vdots \\ a_{N1}(s), \dots, a_{Nj-1}(s); A_N(s) - \sum_{i=1}^N a'_{Ni}(s)x_i^*(z); a_{Nj+1}(s), \dots, a_{NN}(s) \end{vmatrix}$$

† Eqn (7)

$$\Delta_{x_j}(s) = \frac{\begin{vmatrix} a_{11}(s), \dots, a_{1j+1}(s); A_1(s); a_{1j+1}(s), \dots, a_{1N}(s) \\ \vdots \\ a_{N1}(s), \dots, a_{Nj+1}(s); A_N(s); a_{Nj+1}(s), \dots, a_{NN}(s) \end{vmatrix}}{\begin{vmatrix} a_{11}(s), \dots, a_{1j-1}(s); a'_{1k}(s); a_{1j+1}(s), \dots, a_{1N}(s) \\ \vdots \\ a_{N1}(s), \dots, a_{Ni-1}(s); a'_{Nk}(s); a_{Ni+1}(s), \dots, a_{NN}(s) \end{vmatrix}} - \sum_{k=1}^N x_k^*(z)$$

Subjecting (8) to a  $z$  transform and cancelling identical terms, the following relation is obtained:

$$x_j^*(z) \left[ 1 + \left( \frac{\Delta_{x_j^*}^j}{\Delta} \right)^* (z) \right] = \left( \frac{\Delta_A^j}{\Delta} \right)^* (z) - \sum_{k=1}^N x_k^*(z) \left( \frac{\Delta_{x_k^*}^j}{\Delta} \right)^* (z); \quad (k \neq j) \quad (9)$$

Thus the initial system (3) can immediately be transferred into a full system by equations of type (9). The full system of equations of a multiloop sampled-data system has the form:

[illegible]

When writing the determinants forming part of (10), the following symbolization is accepted. The upper index shows which column of the common determinant  $\Delta$  is subject to substitution, while the lower index indicates substitution by coefficients for particular variables. Thus,  $\Delta^b x_j^*$  means that the  $k$ th column of the common determinant  $\Delta$  is to be replaced by coefficients at the  $j$ th discrete coordinate.

System (10) can be solved, relative to the coordinates of interest, by ordinary, algebraic methods. Sampled-data systems with various types of link will now be considered.

### Sampled-data Systems with Continuous Compounding Links

An automatic control system with one pulse element, which can be described by a system of three linear equations with constant coefficients, is studied. The block diagram of the system is given in *Figure 2*, which also shows the transfer functions of both the main loop and the additional links.

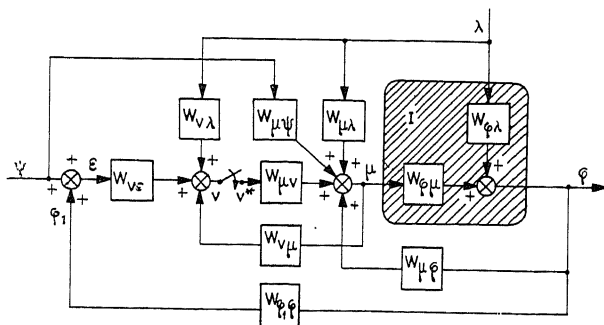


Figure 2. Block diagram of combined control system

The initial system of equations is:

$$\begin{aligned}
& \varphi(s) + 0 - W_{\varphi\lambda}(s) + 0 = W_{\varphi\mu}(s) \lambda(s) \\
& - W_{v\varepsilon}(s) W_{\varphi_1\varphi}(s) \varphi(s) + v(s) - W_{v\mu}(s) \mu(s) + 0 \\
& = W_{v\lambda}(s) \lambda(s) + W_{v\varepsilon}(s) \psi(s) \\
& - W_{\mu\varphi}(s) \varphi(s) + 0 + \mu(s) - W_{\mu v}(s) v^*(z) \\
& = W_{\mu\lambda}(s) \lambda(s) + W_{\mu\psi}(s) \psi(s)
\end{aligned} \tag{11}$$

In accordance with the method expounded above, this system is made into a full one by the deficient equation:

$$\left[1 + \left(\frac{\Delta^2}{\Delta}\right)^*(z)\right] v^*(z) = \left(\frac{\Delta^2}{\Delta}\right)^*(z) \quad (12)$$

Henceforward, only programme and servosystems will be considered; hence, in (11),  $\lambda(s) = 0$ .

From (11) and (12) one can easily find an expression for the controlled coordinate in which one is interested.

$$\varphi(s) = K_2(s)\psi(s) + \frac{K_3(s)}{1 - K_7^*(z) - K_3K_6^*(z)} \{[(K_5 + K_2K_6)\psi]^*(z)\} \quad (13)$$

where

$$\begin{aligned} K_2(s) &= \frac{W_{\varphi\mu}(s) W_{\mu\psi}(s)}{1 - W_{\varphi\mu}(s) W_{\mu\varphi}(s)}; \quad K_3(s) = \frac{W_{\mu\nu}(s) W_{\varphi\mu}(s)}{1 - W_{\varphi\mu}(s) W_{\mu\varphi}(s)}; \\ K_5(s) &= W_{v\varepsilon}(s) + W_{v\mu}(s) W_{\mu\psi}(s); \\ K_6(s) &= W_{\varphi_1\varphi}(s) W_{v\varepsilon}(s) + W_{v\mu}(s) W_{\mu\varphi}(s); \\ K_7(s) &= W_{v\mu}(s) W_{\mu\nu}(s) \end{aligned}$$

*Conditions of Absolute Invariance.* The condition of absolute invariance for servo and programme systems is:

$$\begin{aligned} \varphi(s) &= K_2(s) \psi(s) \\ &+ \frac{K_3(s)}{1 - K_7^*(z) - K_3 K_6^*(z)} \{[(K_5 + K_2 K_6) \psi]^*(z)\} = \psi(s) \\ \text{or} \quad -\varepsilon(s) &= K_2'(s) \psi(s) \\ &+ \frac{K_3(s)}{1 - K_7^*(z) - K_3 K_6^*(z)} \{[(K_5 + K_2 K_6) \psi]^*(z)\} = 0 \end{aligned} \quad \begin{aligned} (14a) \\ (14b) \end{aligned}$$

where  $\varepsilon(s) = \psi(s) - \varphi(s)$  is the system error  $K'_2(s) = K_2(s) - 1$ .

The basic differences between the conditions of invariance for continuous and sampled-data systems is emphasized. While in continuous systems the conditions of absolute invariance do not depend on the form of  $\psi$ , and are determined only by the parameters of the components of the system, in the sampled-data system under consideration, these conditions (14) essentially depend on the form of the input signal  $\psi$ .

It can be shown that the condition of absolute invariance physically signifies the equality to zero of the sum of the individual components of the coordinate  $\varepsilon$  produced both as a result of the direct effect  $\psi$  upon the system, and also on account of the effect *via* the additional (compounding) links.

*Invariance Conditions for Discrete Moments of Time.* The invariance conditions (14) were obtained from the requirement of the equality to zero of coordinate  $\varepsilon$  at any moments of time. One may impose a less rigid requirement—the equality to zero of  $\varepsilon$  at the sampling instants, i.e.,

$$\varepsilon[nT] = 0 \quad (15)$$

The conditions under which (15) is satisfied are called 'conditions of invariance for discrete moments of time'. If (14) is subjected to a  $z$  transform, then it seems that the problem is solved. However, it is easy to show that the invariance conditions for discrete moments of time as well, will depend upon  $\psi$ .

An attempt is made to obtain the conditions, independent of  $\psi$ . Both parts of (14b) are multiplied by

$$\frac{K_5(s) + K_2(s) K_6(s)}{K'_2(s)}$$

and then subjected to a  $z$  transform.

$$-\left(\frac{l}{K'_2}\right)^*(z) = (l\psi)^*(z) + \left(\frac{K_3 l}{K'_2}\right)^*(z) \frac{(l\psi)^*(z)}{1 - K_7^*(z) - K_3 K_6^*(z)} \quad (16)$$

is obtained, where  $l(s) = K_5(s) + K_2(s) K_6(s)$ .

By equating the right-hand side of (16) to zero, the following invariance conditions are obtained for discrete moments of time:

$$1 - K_7^*(z) + \left[\frac{K_3(K_5 + K_6)}{K'_2}\right]^*(z) = 0 \quad (17)$$

The conditions of absolute invariance for a similar continuous system (i.e., a system having the same structure) can be given in the form:

$$1 - K_7(s) + \frac{K_3(s)[K_5(s) + K_6(s)]}{K'_2(s)} = 0 \quad (18)$$

If (18) is subjected to a  $z$  transform, eqn (17) is obtained, i.e., the introduction of a pulse element into an absolutely invariant continuous system does not impair the conditions of invariance for discrete moments of time for the so-called 'fictitious coordinate'

$$\varepsilon_\varphi(s) = \frac{l(s)}{K'_2(s)} \varepsilon(s)$$

As shown by Kremetulo<sup>10</sup> from the equality to zero of  $\varepsilon_\varphi[nT]$ , there still does not follow the equality to zero of  $\varepsilon[nT]$ . The additional conditions will be given, under which  $\varepsilon[nT] = 0$ , and does not depend on the form of  $\psi$ . (14b) is subjected to a  $z$  transform, and then  $1 - K_7^*(z)$ , found from (17), is substituted:

$$-\varepsilon^*(z) = K'_2 \psi^*(z) - \frac{K_3^*(z)}{\left\{\frac{K_3(K_5 + K_6)}{K'_2}\right\}^*(z) + K_3 K_6^*(z)} \times \{K_5 \psi^*(z) + [(K'_2 + 1) K_6 \psi]^*(z)\} \quad (19)$$

The additional condition:

$$\left[\left(\frac{K_5 + K_6}{K'_2} + K_6\right) K'_2 \psi\right]^*(z) = \left(\frac{K_5 + K_6}{K'_2} + K_6\right)^*(z) K'_2 \psi^*(z) \quad (20)$$

Condition (20) is satisfied if  $[(K_5 + K_6)/K'_2] + K_6$  contains proportional components or components with a pure time lag. From (20) and (17) can be found the transfer functions of continuous compounding links.

#### Sampled-data Systems with Discrete Compounding Links

A brief examination will be made of the properties of a typical sampled-data servo-system, the block diagram of which is given in Figure 3. The expression of the system error  $\varepsilon$  is:

$$\varepsilon^*(z) = \frac{1 - W_{\mu\psi}^*(z) W_{\varphi\mu}^*(z)}{1 + W_{v\varepsilon}^*(z) W_{\varphi\mu}^*(z)} \psi^*(z) \quad (21)$$

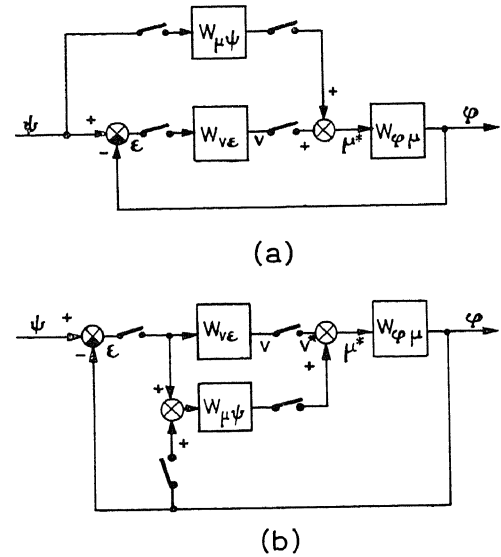


Figure 3. Block diagram of servosystems: (a) with direct link with respect to assignment; (b) with indirect link with respect to assignment

The condition of invariance at discrete moments of time is:

$$W_{\mu\psi}^*(z) = \frac{1}{W_{\varphi\mu}^*(z)} \quad (22)$$

In the general case,  $W_{v\varepsilon}^*(z)$  and  $W_{\varphi\mu}^*(z)$  are the ratio of polynomials according to the positive powers of  $z$ , the power of the numerator being less than that of the denominator. Since  $W_{\mu\psi}^*(z)$  must be inverse to  $W_{\varphi\mu}^*(z)$ , then it cannot be physically realized (advancing components are required for this).

It is important to note that the introduction of the link  $W_{\mu\psi}^*(z)$  and the satisfaction of the invariance condition (22) do not alter the characteristic equation of the system:

$$K_0^*(z) P^*(z) + K_1^*(z) Q^*(z) = 0;$$

$$\left(\frac{K_0^*(z)}{K_1^*(z)} = W_{v\varepsilon}^*(z); \frac{P^*(z)}{Q^*(z)} = W_{\varphi\mu}^*(z)\right) \quad (23)$$

and therefore do not influence the stability of the system.

Examples were given by Kuntsevich<sup>12</sup> to show that even in those cases when  $W_{\mu\psi}^*(z)$ , obtained from condition (22) cannot

be realized, provided it is selected in a particular way, it is possible to increase considerably the accuracy of a sampling servosystem.

When for any reasons it is inconvenient or impossible to introduce the compounding link  $W_{\mu\psi}^*(z)$ , one may introduce into the system additional links, equivalent to the direct compounding link  $W_{\mu\psi}^*(z)$ . Eqn (21) can be brought to the form:

$$\varepsilon^*(z) = \frac{1}{1 + W_{v\varepsilon}^*(z) W_{\varphi\mu}^*(z)} \psi^*(z) - \frac{W_{\mu\psi}^*(z) W_{\varphi\mu}^*(z)}{1 + W_{v\varepsilon}^*(z) W_{\varphi\mu}^*(z)} [\varepsilon^*(z) + \varphi^*(z)] \quad (24)$$

It is not difficult to see that (24) is met by the diagram shown in Figure 3 (b).

If (22) is satisfied, then the condition of absolute invariance has the form:

$$\frac{\psi(s)}{\psi^*(z)} = \frac{W_{\varphi\mu}(s)}{W_{\varphi\mu}^*(z)} \quad (25)$$

The latter equality can be satisfied only in some particular cases, and, as shown by Krementulo<sup>11</sup>, requires the inclusion of advancing components if  $\psi[0] = 0$ .

#### Sampled-data Systems With Pulse-continuous Compounding Links

In this section a servosystem will be used as an example to show that when pulse-continuous links are used it is in principle possible to achieve absolute invariance in a combined sampled-data system.

Assume that the block diagram is predetermined, i. e.,  $W_{v\varepsilon}(s)$ ,  $W_{\varphi\mu}(s)$  and  $W_{\varepsilon\varphi}(s)$  are known. A compounding link with respect to the input signal  $\psi$   $W_{\mu\psi}(s)$  is introduced to improve the dynamic properties. The transfer function of this link has to be determined.

The expression for the system error is:

$$-\varepsilon(s) = [W_{\mu\psi}(s) W_{\varphi\mu}(s) - 1] \psi(s) + \frac{W_{v\varepsilon}(s) W_{\varphi\mu}(s)}{1 + W_{v\varepsilon}(s) W_{\varphi\mu}(s) W_{\varepsilon\varphi}^*(z)} [\psi^*(z) + W_{\mu\psi} W_{\varphi\mu} W_{\varepsilon\varphi} \psi^*(z)] \quad (26)$$

Having equated  $\varepsilon(s)$  to zero the condition of invariance of the system is obtained from which the transfer function of the compounding link can be determined:

$$W_{\mu\psi}(s) = \frac{1}{W_{\varphi\mu}(s)} + \frac{W_{v\varepsilon}(s)}{\psi(s)} [W_{\varepsilon\varphi} \psi^*(z) - \psi^*(z)] \quad (27)$$

The signal of the compounding link  $v_1(s)$  equals:

$$v_1(s) = \frac{\psi(s)}{W_{\varphi\mu}(s)} + W_{v\varepsilon}(s) [W_{\varepsilon\varphi} \psi^*(z) - \psi^*(z)] \quad (28)$$

This signal can be realized with the aid of the scheme shown in Figure 4 (b). In a similar continuous system, the compounding link with respect to  $\psi$ , chosen from the conditions of absolute invariance, equals:

$$W_{\mu\psi}(s) = \frac{1}{W_{\varphi\mu}(s)} + W_{v\varepsilon}(s) [W_{\varepsilon\varphi}(s) - 1] \quad (29)$$

It can be seen that for both the sampled-data and the continuous system the compounding link has one and the same structure and consists of identical components. The difference lies in the fact that in an absolutely invariant sampled-data system some of the components are connected *via* additional pulse elements operating synchronously and in phase with the main one. What has already been said also holds in the case when real pulse elements are used.

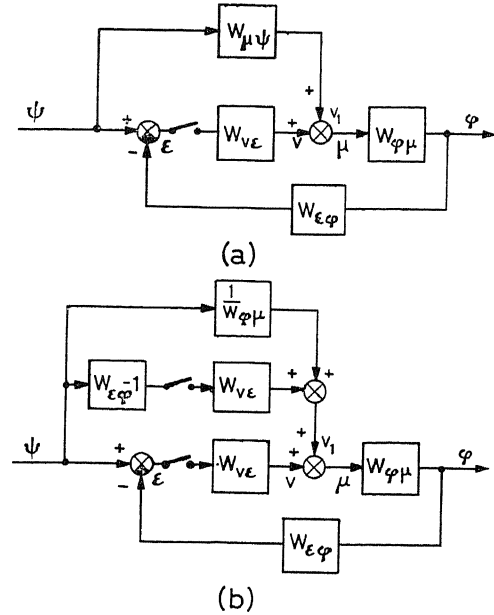


Figure 4. Block diagram of structural scheme of combined servosystem

#### Extremal Sampled-data Systems

##### Systems without Compounding Links

Today a large number of extremal sampled-data systems of various types are known, which have been studied by many scientists. But certain specific features of these systems remain unexplained. An analysis will be made of the known extremal sampled-data systems on the basis of full and precise equations of dynamics of only one system which, as was shown in (29), provides the best tracking quality with continuous drift of the extremum, and whose properties are at the same time closest to those of a hypothetical system measuring the position of the extremum point without any errors.

As in most works, the controlled plant with extremal characteristics will be considered to be one which consists of a linear inertial component and an inertia-less component with extremal characteristics.

The equation of the non-linear component, taking into account the action of two kinds of disturbances (or two components of one and the same disturbance), which displace the extremum point, will be written in the form:

$$\varphi = -\alpha_3 (x + \psi)^2 + \lambda \quad (30)$$

where  $\varphi$  is the index of the extremum, and  $\psi$ ,  $\lambda$  are disturbances of an arbitrary kind, inaccessible to direct measurement by virtue of the conditions of the problem. Let the remaining



equations of the extremal system\* (see *Figure 5*) in the absence of the components shown in *Figure 5* by the dotted line, be:

$$x(s) = W_{xM}(s)M(s) \quad (31)$$

where

$$M = \mu + m \quad (31a)$$

$$m(s) = W_{mv}(s) V^*(z) \quad (32)$$

where

$$V^*(z) = a'_M \frac{z}{1+z} \quad (32a)$$

$$\mu(s) = W_{\mu u}(s) u^*(z) \quad (33)$$

$$y_n = \Delta \varphi_{n-1} (-1)^n \quad (34)$$

$$u^*(z) = W_{uy}^*(z) y^*(z) \quad (35)$$

The error of the system is denoted as:

$$e = \mu' + \psi \quad (36)$$

and also the notations are introduced

$$x^*(z) = \mu'^*(z) + m'^*(z) \quad (37)$$

where

$$\mu'^*(z) = W_{xM} W_{\mu u}^*(z) u^*(z) \quad (37a)$$

$$m'^*(z) = W_{mv} W_{\psi M}^*(z) V^*(z) \quad (37b)$$

On the basis of (37b) and (32, 32a), the modulating effect  $m'_n$ , scaled to the input of the non-linear element, can be represented in the form

$$m'_n = a_M \cos \pi n = a_M (-1)^n \quad (38)$$

where  $a_M$  is determined from the particular solution of the difference equation

$$a_M(-1)^n = a'_M W_{xM} W_{mv}(E)(-1)^n \quad (39)$$

which is obtained following the replacement of (32) by the corresponding difference equation.

Solving jointly (30), (36), (37) and (38) gives

$$y_n = -2a_M\alpha_3(e_n + e_{n-1}) + \Delta\lambda_{n-1}(-1)^n - \alpha_3(e_n^2 - e_{n-1}^2)(-1)^n \quad (40)$$

From (40) it can be seen that the signal on the output of the component (34), apart from the useful component proportional to the error, also contains additional terms, one of which  $\Delta\lambda_{n-1}(-1)^n$  reflects the influence of the disturbance  $\lambda_n$ , and the third term shows that the measurement of the position of the system relative to the extremum point is not ideal.

Further replacing (35)–(37) by their corresponding difference equations, and solving them jointly with (40) and (37a), the equation of the dynamics of the system is obtained in the form of a non-linear difference equation with time-varying coefficients

$$[2 a_M \alpha_3 W(E)(E+1)+E] e_n - \alpha_3 W(E) [e_{n+1}^2 - e_n^2] \cos \pi n]$$

$$= \psi_{n+1} - W(E) [\Delta \lambda_n \cos \pi n]$$

$$\text{where } W(E) = W_{xM} W_{uu}(E) W_{uv}(E) \quad (41)$$

As was shown by Kuntsevich<sup>29, 30</sup>, the non-linear eqn (41) has the peculiarity that for a particular correlation between the system parameters and the speed of variation of disturbances  $\psi_n, \lambda_n$  the stability of the system is impaired, whereas analysis of the linearized equation obtained from (41), disregarding the non-linear terms (as done by Chang<sup>25</sup>, Van-Neis<sup>26</sup> and Ivakhnenko<sup>27</sup>) does not permit one to detect this phenomenon. Therefore the feasibility of constructing an adaptive system, the error of which would be invariant in relation to  $\psi_n, \lambda_n$ , acquires particular interest, since it involves not only the improvement of the quality of the system, but also the increasing of its stability margin.

### *Invariance of Extremal Control Systems with Indirect Compounding Links*

Since, by virtue of the conditions of the problem, the possibility of direct measurement of the signals  $\psi$  and  $\lambda$  is excluded, the possibility will be considered of using indirect compounding links with respect to  $\psi$  and  $\lambda$  similar to those considered above.

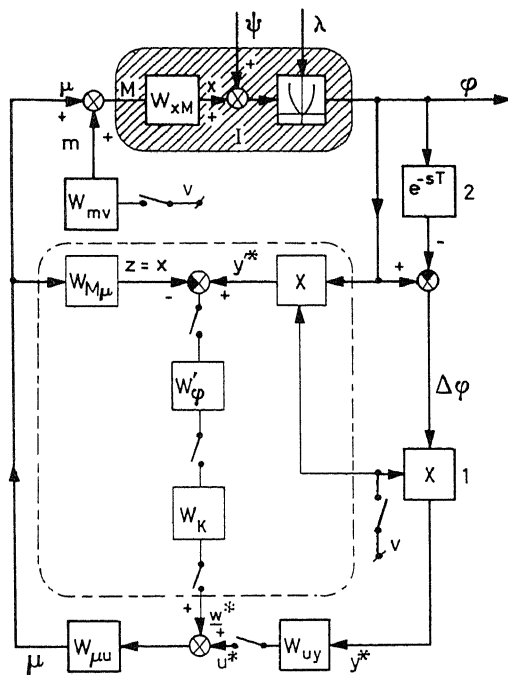


Figure 5. Block diagram of a difference-type sampled-data extremal system with indirect compounding link

I: plant; 1: multiplying unit; 2: memory element

Here (31) is the equation of the linear part of the plant, (32) the equation of the modulation circuit, (34) the equation of a controller with synchronous detector, (35) the equation of the correcting elements, (33) the equation of the servomotor and  $x, \mu, \varphi, u$  and  $y$  the controlled coordinates.

Henceforward it is taken that the dynamic properties of the plant and the slope  $\alpha_3$  of the extremal characteristic are constant or quasi-constant.

\* Since the system under review is non-linear, then strictly speaking, neither the ordinary nor the discrete Laplace transform is applicable to it. Therefore the final results will be obtained with the aid of a set of non-linear difference equations. To simplify things, the Laplace transform will only be used in application to the linear components.

Consideration will first be given to the possibility of attaining invariance of system error at discrete moments of time, relative to  $\psi_n^*$ .

From (36), (37), and (37a) and also from Figure 5, it follows that

$$\psi^*(z) = e^*(z) - \mu^*(z) \quad (42)$$

or

$$\psi^*(z) = e^*(z) - W_{xM} \mu^*(z) \quad (42a)$$

For the construction of the correcting link with respect to  $\psi$  in accordance with (42a), the variable  $\mu'_n$  can be obtained with the aid of a model of the linear part of the controlled plant (see Figure 5\*). A signal proportional to  $e_n$  (or, more strictly, containing  $e_n$ ) can be obtained at the output of an additional synchronous detector (see the part of Figure 5 outlined by a broken line), the equation of which is:

$$y'_n = \varphi_n (-1)^n \quad (43)$$

Solving (30), (36) and (43) jointly, gives

$$y'_n = -2a_M \alpha_3 e_n - \alpha_3 (e_n^2 + a_M^2) (-1)^n + \lambda_n (-1)^n \quad (44)$$

For filtering the parasitic quasi-periodic terms of signal (44) at the output of the detector in the network in Figure 5, a low-frequency filter is provided.

Taking this into account, the signal on the output of the additional control loop is written in the form

$$w_n \approx D(E) \psi_n \quad (45)$$

where

$$D(E) = -2a_M \alpha_3 W'_\varphi(E) W_K(E)$$

Omitting the intermediate operations, the equation of the dynamics of the system in Figure 5, with an additional control loop, is obtained, on the basis of the equations cited above and also eqn (45), in the form

$$\begin{aligned} & [2a_M \alpha_3 W(E)(E+1) + E] e_n - \alpha_3 W(E) [(e_{n+1}^2 - e_n^2) \cos \pi n] \\ & = [1 - 2a_M \alpha_3 W_{xM} W_{\mu\mu}(E) W'_\varphi(E) W_K(E)] \psi_{n+1} \\ & - W(E) \Delta \lambda_n \cos \pi n \end{aligned} \quad (46)$$

By equating to zero the operator comultiplier for  $\psi$  in the right-hand side of (46), an expression is obtained of the impulse transfer function  $W_K(E)$ , which ensures the invariance of the system from  $\psi_n$  at discrete moments of time

$$W_K(E) = \frac{1}{2a_M \alpha_3} \frac{1}{W'_\varphi(E) W(E)} \quad (47)$$

From (46) it can be seen that the satisfaction of the conditions of invariance (47), and the presence of the filter in the compounding-link network (as distinct from the filter in the main network of the controller), do not alter the form and coefficients of the left-hand side of the equation of the dynamics of the system, i.e., do not directly influence the stability of the system.

When the required transfer function  $W_K^*(z)$  is physically unrealizable, then, as for ordinary servosystems, a considerable improvement of accuracy (increasing the degree of astatism) can be achieved by appropriate selection of the transfer function

\* It is noted that in contrast to ordinary servosystems, in which the input signal may also contain a noise which has to be suppressed as effectively as possible, the task of an extremal system in all cases is the complete performance of signal  $\psi$ .

$W_K^*(z)$ . An example is given in the Appendix of the method of selection of the coefficients of the transfer function  $W_K^*(z)$ .

In deriving the conditions of invariance (47), the quasi-periodic non-linear terms in (44) were disregarded in order to simplify the investigation. As follows from the example in the Appendix (see also Figure 6), the influence of these terms is in fact small.\*

A brief examination will now be made of the possibility of minimization (or complete elimination) of the system error due to  $\lambda$ . From the equation of the system dynamics (46) and (40), it follows that for the predetermined structure the possibility of constructing a correcting link with respect to  $\lambda(t)$  in a

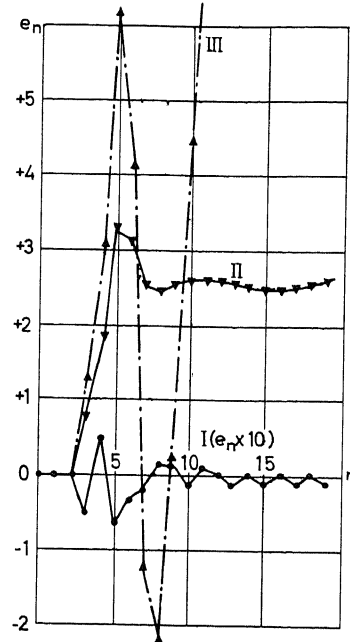


Figure 6. Transients of extremal system for  $\psi_n = \beta n$ ,  $\Delta \lambda_n = 0$

I: in system (54) with compounding link for satisfaction of condition (50); ( $\alpha_1 \alpha_2 = 0.4$ ;  $\alpha_3 = 1$ ;  $d_1 = 0.4$ ;  $d_2 = 0.8$ ;  $\beta = 3.5$ )

II: in system (55) (ditto, but  $d_1 = d_2 = 0.4$ ;  $\beta = 2$ )

III: in system (55) (ditto, but for  $\beta = 3.5$ )

similar way as with respect to  $\psi$ , without constructing an analogue of the non-linear component, is excluded. By virtue of this, with the structure adopted, only methods of minimizing the influence of  $\lambda(t)$  can be considered. One such method, based on the selection of the corresponding function  $W_{\mu\mu}^*(z)$  was considered by Chang<sup>25</sup>, Van-Neis<sup>26</sup> and Ivakhnenko<sup>27</sup>. The results obtained by Tou<sup>24</sup> may also be used here.

## Appendix

Example—In Figure 5 let

$$W_{xM}(s) = \frac{\alpha_1}{\tau_1 s + 1}; \quad W_{\mu\mu}(s) = \frac{\alpha_2}{s}$$

\* The system in Figure 5 was checked experimentally on an electronic analogue by A. A. Tunik, and the check confirmed the effectiveness of the introduction of indirect correction<sup>31</sup>.

to which there corresponds

$$W_{xM} W_{uu}^*(z) = \frac{\alpha_1 \alpha_2 (1-d_1) z}{(z-1)(z-d_1)}$$

and further, let

$$W_{uy}^*(z) = \frac{B_1^*(z)}{B_2^*(z)}$$

where  $B_1^*(z)$  and  $B_2^*(z)$  are polynomials of  $z$ ,  $d_1 = e^{-T/\tau_1}$ .

It will be taken that

$$W_\varphi'(s) = \frac{1 - e^{-sT}}{s} \frac{1}{\tau_2 s + 1}$$

to which there corresponds

$$W_\varphi'(z) = \frac{1-d_2}{z-d_2}, \quad (d_2 = e^{-T/\tau_2})$$

It is not difficult to see that in the given case the impulse transfer function  $W_K^*(z)$ , as determined from (47), which is required for attainment of the conditions of invariance, is physically unrealizable, and only the approximate satisfaction of the conditions of invariance can be considered; by virtue of this,  $W_K^*(z)$  will be sought in the form of the series

$$W_K^*(z) = \sum_{i=1}^K C_i \left( \frac{z-1}{z} \right)^i \quad (48)$$

Denoting the left-hand side of equation (46) by  $L(E)e_n$  in order to abbreviate the notation, one can write it for  $\Delta \lambda_n = 0$  in the given example, bearing in mind (48), in the form:

$$\begin{aligned} L(E)e_n = & EB_2(E) \{ -2a_M \alpha_1 \alpha_2 \alpha_3 (1-d_1)(1-d_2) \\ & \times [C_1 \Delta \psi_n + C_2 \Delta^2 \psi_{n-1} + \dots + C_K \Delta^K \psi_{n-K+1}] \\ & + \Delta^3 \psi_n + [(1-d_1) + (1-d_2)] \Delta^2 \psi_n + (1-d_1)(1-d_2) \} \Delta \psi_n \end{aligned} \quad (49)$$

Provided

$$C_1 = \frac{1}{2a_M \alpha_1 \alpha_2 \alpha_3} \quad (50)$$

the error from the first difference  $\psi_n$  is eliminated, since, when this is satisfied, the equation of the system adopts the form

$$\begin{aligned} L(E)e_n = & EB_2(E) \{ -2a_M \alpha_1 \alpha_2 \alpha_3 (1-d_1)(1-d_2) [C_2 \Delta^2 \psi_{n-1} + \dots \\ & + C_K \Delta^K \psi_{n-K+1}] + \Delta^3 \psi_n + (2-d_1-d_2) \} \Delta^2 \psi_n \end{aligned} \quad (51)$$

Further taking

$$C_2 = \frac{(2-d_1-d_2)}{2a_M \alpha_1 \alpha_2 \alpha_3 (1-d_1)(1-d_2)} \quad (52)$$

and bearing in mind that

$$\Delta^i \psi_n - \Delta^i \psi_{n-1} = \Delta^{i+1} \psi_{n-1}$$

(51) can be rewritten in the form

$$\begin{aligned} L(E)e_n = & EB_2(E) \{ -2a_M \alpha_1 \alpha_2 \alpha_3 (1-d_1)(1-d_2) [C_3 \Delta^3 \psi_{n-2} + \dots \\ & + C_K \Delta^K \psi_{n-K+1}] + \Delta^3 \psi_n - C_2 \} \Delta^3 \psi_{n-1} \end{aligned} \quad (53)$$

from which it will be seen that, irrespective of the coefficients  $W_{uy}^*(z)$  the error is eliminated from the second difference  $\psi_n$ . Since further increasing of the degree of astatism on account

of the correcting link is impossible in the given example,  $C_i = 0$  will be taken for  $i \geq 3$ .

For quantitative evaluation of the quasi-periodic terms in (46), which have not been taken into account, in Figure 6 the transient in an extremal system is plotted, taking into account these terms for  $\psi_n = \beta n$ ,  $\Delta \lambda_n = 0$  for eqn (46).

For the transfer function of the components cited in the example under consideration and for  $W_{uu}^* z = 1$ , the precise equation of the dynamics of the system has the form:

$$\begin{aligned} & A_0 e_{n+3} + A_1 e_{n+2} + A_2 e_{n+1} + A_3 e_n \\ & - \alpha_\Sigma (1-d_1) [e_{n+2}^2 - e_{n+1}^2 + d_2 (e_{n+1}^2 - e_n^2)] (-1)^n \\ & - \alpha_\Sigma (1-d_1)(1-d_2) [e_{n+1}^2 + e_n^2 + 2a_M^2] (-1)^n = 0 \end{aligned} \quad (54)$$

where

$$\begin{aligned} A_0 &= 1; \quad A_1 = 2a_M \alpha_\Sigma (1-d_1) - (1+d_1+d_2); \\ A_2 &= 2a_M \alpha_\Sigma (1-d_1)(1-d_2) + d_1 + d_2 + d_1 d_2; \\ A_3 &= -d_1 d_2 - 2a_M \alpha_\Sigma (1-d_1) d_2; \quad \alpha_\Sigma = \alpha_1 \alpha_2 \alpha_3 \end{aligned}$$

Here, for comparison, the transient processes in an extremal system without correcting link with respect to  $\psi_n$  have been plotted, in which  $W_{xM}(s)$  and  $W_{uu}(s)$  are the same as given above, and a low-frequency filter with transfer function

$$W_\varphi(s) = \frac{1 - e^{-sT}}{s} \frac{1}{\tau_3 s + 1}$$

is included into the main extremal-control network.  $Tz$  transform of  $W_\varphi(s)$  is

$$W_\varphi^*(z) = \frac{1-d_3}{z-d_3}$$

where  $d_3 = e^{-T/\tau_3}$ .

Bearing this remark in mind, for the given case, the equation of the dynamics (41) of the system adopts the form

$$\begin{aligned} & A'_0 e_{n+3} + A'_1 e_{n+2} + A'_2 e_{n+1} + A'_3 e_n \\ & - \alpha_\Sigma (1-d_1) [e_{n+2}^2 - e_{n+1}^2 + d_3 (e_{n+1}^2 - e_n^2)] (-1)^n \\ & = \Delta^3 \psi_{n+1} + [(1-d_1) + (1-d_3)] \Delta^2 \psi_{n+1} \\ & + (1-d_1)(1-d_3) \Delta \psi_{n+1} \end{aligned} \quad (55)$$

where

$$\begin{aligned} A'_0 &= 1; \quad A'_1 = 2a_M \alpha_\Sigma (1-d_1) - (1+d_1+d_3); \\ A'_2 &= 2a_M \alpha_\Sigma (1-d_1)(1-d_3) + d_1 + d_3 + d_1 d_3; \\ A'_3 &= -2a_M \alpha_\Sigma (1-d_1) d_3 - d_1 d_3; \quad \alpha_\Sigma = \alpha_1 \alpha_2 \alpha_3 \end{aligned}$$

As can be seen from the curves in Figure 6, an increase in  $\beta$  (the rate of drift of the extremum) leads to the loss of the stability of the system (55). Thus the introduction of compounding links with respect to  $\psi_n$  not only improves the quality of the system, but also preserves its stability, thus extending the sphere of application of extremal systems to the case of high extremum drift rates.

## References

- SCHIPANOV, G. V. Theory and method of design of automatic controllers. *Automat. Telemekh., Moscow* 1 (1939)
- KULEBAKIN, V. S. The theory of invariance of regulating and control systems. *Automatic and Remote Control*. p. 106. 1961. London; Butterworths

- <sup>3</sup> PETROV, B. N. The invariance principle and the conditions for its application during the calculation in the design of linear and non-linear systems. *Automatic and Remote Control*. p. 117. 1961. London; Butterworths
- <sup>4</sup> IVAKHENKO, O. G. *Automatika* (1961)
- <sup>5</sup> KOSTYUK, O. M. *Automatika* 1 (1961)
- <sup>6</sup> BELYA, K. K. The invariance of the controlled magnitude of an automatic device from certain of its parameters. *Izv. Akad. Nauk SSSR, Otdel Tekhn. Nauk, Energ. Automat.* 6 (1961)
- <sup>7</sup> Invariance theory and its application in automatic devices. *Trud. Soveshch. Sostoyavshegosya v g. Kiev*, 16-20 sent., 1958 (Proc. of a meeting held in Kiev, Sept. 16-20, 1958), Moscow, 1959
- <sup>8</sup> TSYPKIN, YA. Z. *Automatika* 1 (1958)
- <sup>9</sup> TOU, J. Digital compensation for control and simulation. *Proc. Inst. Radio Engrs, N.Y.* Vol. 45, No. 9 (1957)
- <sup>10</sup> KREMENTULO, YU. V. *Automatika* 1 (1962)
- <sup>11</sup> KREMENTULO, YU. V. *Automatika* 2 (1960)
- <sup>12</sup> KUNTSEVICH, V. M. *Automatika* 1 (1962)
- <sup>13</sup> GRISHCHENKO, L. Z., and BOLDYREVA, D. F. The invariance of automatic sampled-data control systems. *Automatika* 2 (1962)
- <sup>14</sup> STREITZ, V., and RUZHICHKA, I. The theory of autonomy and invariance of multiparameter control systems with digital controllers. *Izv. Akad. Nauk SSSR, Otdel Tekhn. Nauk. Energ. Automat.* 5 (1961)
- <sup>15</sup> FEDOROV, S. M. Delay in the synthesis of servosystems with digital computers. *Izv. Akad. Nauk SSSR, Otdel Tekhn. Nauk. Energ. Automat.* 4 (1961)
- <sup>16</sup> TSYPKIN, YA. Z. *Teoriya Impul'snykh Sistem* (Theory of sampled-data systems) 1958. Moscow; Fizmatgiz
- <sup>17</sup> BURSHTIN, I. M. Solving equations of multiloop sampled-data systems. *Automat. Telemekh., Moscow* 12 (1961)
- <sup>18</sup> RAGAZZINI, J. R., and FRANKLIN, G. F. *Sampled-data Control Systems*. 1958. New York; McGraw-Hill
- <sup>19</sup> JURY, E. J. *Sampled-data Control Systems*. New York; John Wiley. London; Chapman and Hall
- <sup>20</sup> LENDARIS, G. G. and JURY, E. J. *Input-output Relationships for Multisampled-loop Systems Applications and Industry*. Jan. 1960
- <sup>21</sup> TOU, J. A simplified technique for determination of output transforms of multiloop multisampler variable-rate discrete-data systems. *Proc. Inst. Radio Engrs, N. Y.* 49, 3
- <sup>22</sup> TOU, J. *Digital and Sampled-data Control Systems*. 1959. New York; McGraw-Hill
- <sup>23</sup> SALZER, G. M. Signal-flow reduction in sampled-data systems. *Wescon Conventional Record, Inst. Radio Engrs, N. Y.* Pt IV (1957)
- <sup>24</sup> TOU, J. Statistical design of linear discrete-data control systems via the modified z-transform method. *J. Franklin Inst.* 271, 4 (1961)
- <sup>25</sup> CHANG, S. S. L. Optimization of the adaptive function by the z-transform method. *A.I.E.E. Conf. Pap. NCP 59-1296* (see also *Synthesis of Optimum Control Systems*. Ch. 10, 11. 1961. New York; McGraw-Hill)
- <sup>26</sup> VAN-NEIS, R. I. *Automatika* 1, 2 (1961)
- <sup>27</sup> IVAKHENKO, A. G. Comparison of cybernetic extremal sampled-data systems characterized by extremum search strategy. *Automatika* 3 (1961)
- <sup>28</sup> FELDBAUM, A. A. *Vychislitelnye Ustroistva v Avtomatika* (Computers and Automation) 1959. Moscow
- <sup>29</sup> KUNTSEVICH, V. M. A study of sampled-data extremal systems with extremum drift. *Automat. Telemekh., Moscow* 7 (1962)
- <sup>30</sup> KUNTSEVICH, V. M. Invariance of sampled-data extremal systems without disturbance links. *Automatika* 3 (1962)
- <sup>31</sup> TUNIK, A. A. *Automatika* 6 (1962)

## DISCUSSION

G. AXELBY, *Westinghouse Electric Corporation, Linthicum, Md., U.S.A.*

Would the author please comment on the possible application of the system described in his paper.

YU. V. KREMENTULO, *in reply*

The method of improving the quality of sampled-data systems, which is based on the compounding link, can, of course, be practically realized if we can build the model of the plant. This is possible in several cases, and I can give two examples. The first relates to the problem of measurement of the rate of rolling by means of correlation. The second relates to the problem of water refining. In this case the dynamic model of the plant permits the use of the extremum controller without test signals<sup>1</sup>.

## Reference

- <sup>1</sup> VASILJEV, V. I. *Automatika* (1962)

E. JURY, *University of California, Berkeley, California, U.S.A.*

This paper presents a general and thorough study of the invariance problem in sampled-data systems, based on the assumption expounded by the authors. I would like to mention two points.

(1) The introduction of the indirect compounding link is a useful artifice to improve on the invariance of the system. However, this can also be achieved by introducing additional components as well as with a direct link. Does the indirect compounding link represent the minimum number of components or could it be further simplified without much affecting the system operation?

(2) I would mention that by varying the sampling rate as a function

of the error or other signals of the system, one could also achieve the objectives introduced by the authors. I have studied some work along these lines in connection with pulse-frequency modulated feedback systems where certain favourable results have been achieved. A short discussion<sup>1</sup> of PFM (pulse frequency modulated) systems has been published and an extensive discussion of such problems has been recently presented.

## Reference

- <sup>1</sup> JURY, E. I. On sampling aspects in feedback control systems, *Regelungstechnik*, March, 1963

YU. V. KREMENTULO, *in reply*

The problem which deals with the minimum number of components used for indirect measurement of disturbances is complex and it is difficult to give an answer in the general case. It depends on the concrete task and concrete control system; but usually indirect measurement of disturbances leads to a smaller number of components. We must use the methods of indirect measurement of disturbances because in many cases the causes of the disturbances are unknown, for example, in the extremum systems.

I completely agree with the second remark. Of course, we have many methods of improving performance of sampled-data systems, but many of these methods have not been completely studied. Examples of such methods are: varying the sampling rate, changing the shape of pulses, etc. In this connection, I mention the interesting paper which was presented by Dr. B. Friedland at the 1st I.F.A.C. Congress. This paper deals particularly with an amplifier which has a periodically varying gain to improve a transient performance of the sampled-data servo system.

# Optimization and Invariance in Control Systems with Constant and Variable Structure

B. N. PETROV, G. M. ULANOV and S. V. EMEL'YANOV

## Summary

The paper considers invariance in the automatic control systems subject to external disturbances with known statistics. Invariance conditions established by means of  $K(D)$  transform theory are generalized for systems with known disturbance statistics.

A new design principle for the systems invariant to any continuous control function is developed. Usually error signal is independent of control action only if the right-hand side of the system's non-homogeneous differential equation vanishes. To meet this requirement one must use control action derivatives in the control algorithm. In this case any variation of system parameters would make these invariance conditions invalid.

It is shown that with the aid of open-loop variable structure systems one can perform a wide range of control functions without static errors. In this case there is no need to meet this classical invariance condition. Note that combined servo systems with variable open-loop structure are insensitive to certain variations of the system parameters.

## Sommaire

On traite de l'invariance des systèmes de commande automatique assujettis à des perturbations extérieures à caractère statistique connu. Les conditions d'invariance établies à l'aide de la théorie de la transformée  $K(D)$  sont généralisées à ces systèmes.

On développe un nouveau principe de calcul de l'invariance des systèmes à l'égard de toute fonction de commande continue. Généralement, le signal d'erreur n'est indépendant de l'action de commande que si la partie droite de l'équation différentielle non-homogène du système disparaît. Pour satisfaire à cette exigence, on doit utiliser dans l'algorithme de commande les dérivées de l'action de commande. Dans ce cas, toute variation des paramètres du système invaliderait ces conditions d'invariance.

On montre qu'avec des systèmes en chaîne ouverte et à structure variable, on peut obtenir une large gamme de fonctions de commande sans erreurs statiques. On évite ainsi de satisfaire à la condition d'invariance classique. Notons d'ailleurs que ce genre de systèmes asservis combinés avec des chaînes ouvertes à structure variable sont insensibles à certaines variations des paramètres du système.

## Zusammenfassung

Dieser Beitrag betrachtet die Invarianz von Regelungen, die äußeren Störeinflüssen mit bekannter statistischer Verteilung unterliegen. Invarianzbedingungen, die auf Grund der Theorie der  $K(D)$  Transformation aufgestellt wurden, werden für Systeme mit bekannter statistischer Störung verallgemeinert.

Ein neues Bauprinzip für Systeme, die für beliebige, stetige Regel-funktionen invariant sind, wird entwickelt. Die Regelabweichung ist gewöhnlich nur dann von der Regelwirkung unabhängig, wenn die rechte Seite der das System beschreibenden nicht-homogenen Differentialgleichung verschwindet. Zu diesem Zweck braucht man Ableitungen der Regelwirkung im gegebenen Regelungsalgorithmus. In diesem Falle würde eine beliebige Änderung der Systemparameter diese Invarianzbedingungen ungültig werden lassen.

Es wird gezeigt, daß mit Hilfe von veränderlichen Steuerungsstrukturen viele Regelfunktionen ohne statischen Fehler beherrscht

werden können. In diesem Falle ist es nicht notwendig, die klassische Invarianzbedingung zu erfüllen. Dabei ist zu beachten, daß kombinierte Servosysteme mit veränderlichen Steuerungsstrukturen auf gewisse Änderungen der Systemparameter nicht ansprechen.

## Invariance and Optimization in Automatic Control Systems

### Optimization of Automatic Control Systems and $K(D)$ Image Theory

The object of the general theory of optimization of automatic control systems with respect to accuracy is the optimal synthesis of control systems operating under conditions of continuously-acting disturbances.

In the deterministic set-up of the problem<sup>1-3, 7, 8</sup> the optimality criterion is the achievement of the highest degree of accuracy of the automatic control system, as measured by the error  $\varepsilon$ , which is equal to the difference between the desired  $g(t)$  and the realized  $x(t)$  value of the state of the system  $\varepsilon = g(t) - x(t)$ . In the case of statistical synthesis the optimal system found from the probability characteristics of the controlling signal and the interference, has a transfer function  $\Phi_{opt}$ , and possesses the greatest accuracy only in the mean.

The main results relating to the construction of optimal systems in the case of the deterministic set-up, have been obtained by the theory of invariance, on the basis of which there can be effected the construction of automatic control systems with an error  $\varepsilon$ , equal to zero or extremely small in the presence of disturbances, the measurement or use of which for the purposes of control is feasible. The conditions of the theory of invariance of automatic control systems, in the case when disturbance links do not nullify the numerator of the transfer function (and thus the corresponding transfer function), and when  $f(t)$  is specified, are expressed with the aid of the  $K(D)$  image introduced by Kulebakin

$$K(D) \cdot f(t) = 0, \quad K(D) \neq 0, \quad f(t) \neq 0 \dots \quad (1)$$

$K(D)$  and  $f(t)$  are linked by the conditions of the operator  $K(D)$  image of the functions<sup>1</sup>. In this case for a stable system its transfer function must either be the conform  $K(D)$  image or have this operator  $K(D)$  image as co-multiplier.

In the statistical set-up, with regard to the determination of the transfer function of a control system in the case when it has an infinite memory, according to the mean-square error minimum criterion, one of the main results was obtained by Wiener. Obviously, in one case it is possible to establish precisely the correspondence of optimal systems in the case of the statistical and deterministic set-up of the problem. When the dispersion  $f(t)$  tends to zero, Wiener's optimal system and the optimal system

as determined by the conditions of invariance coincide and should, strictly speaking, lead to the same results. The generality of systems obtained in this case according to Wiener, and of invariant systems, in particular systems meeting the condition of Kulebakin's  $K(D)$  image, are demonstrated. Taking the interval of observation of  $f(t)$  to be infinite, and thus being concerned only with the forced output of the system, the transfer function of a Wiener optimal system is characterized by the magnitude of the  $MS$  error  $\bar{\varepsilon}^2$  (ref. 6):

$$\bar{\varepsilon}^2 = \frac{1}{2\pi} \int_{-\infty}^{\infty} \{S_n(\omega) - |\Phi_{opt}(j\omega)|^2 S_f(\omega)\} d\omega \dots \quad (2)$$

$S_f(\omega)$  is the spectral density of  $f(t)$ ,  $S_n(\omega)$  the spectral density of the desired output signal. In the reviewed problems of control for stabilization  $S_n(\omega)$  is conformally equal to zero, since, with complete filtering of external disturbance  $f(t)$ , the desired output of the system must be conformally equal to zero. The conditions of zero error  $\bar{\varepsilon}_{min}^2 = 0$  lead to the following requirement in respect of the optimal transfer function of an automatic control system:

$$\bar{\varepsilon}^2 = 0 \quad S_n(\omega) = 0 \quad (3)$$

$$|\Phi_{opt}(j\omega)|^2 S_f(\omega) = 0 \quad (4)$$

The latter can be satisfied for  $\Phi(p) \cdot f(t) = 0$ , which is a sufficient condition.

In the case indicated, when

$$\Phi(p) = \frac{\Delta_1(p)}{\Delta(p)} = 0$$

where  $\Delta_1(p)$  is the numerator of the transfer function, and  $\Delta(p)$  is the characteristic polynomial of the automatic control system, expression (4) can be found for (a)  $\Delta(p) = 0$  or (b)  $K(p) \rightarrow \infty$ , where  $K(p)$  is the transfer coefficient of the automatic control system.

The above-mentioned conditions correspond to the known conditions of invariance, the realization of which in physical systems is determined specially.

Without individually examining the above-mentioned possibilities (for  $\Phi(p) = 0$ ), the case of the non-zero operator  $\Phi(p) \neq 0$  will be considered.

If  $\Phi_{opt} \neq 0$  and  $S_f \neq 0$  the satisfaction of condition (4) is possible when

$$\Phi_{opt}(p) \cdot f(t) = 0 \quad (5)$$

This requirement corresponds to the condition of invariance optimal according to Wiener in respect of disturbance  $f(t)$ , and coincides with the  $K(D)$  image<sup>1</sup>. An analogous method is used to establish the community of invariant systems and systems optimal according to Wiener, in the case of other control problems. Thus the  $K(D)$  image can serve as a tool for automatic control systems optimization theory.

As an example, consideration is given to the forced motion of an automatic control system under the influence of an external disturbance, which is described by the equation

$$\Delta(p) \cdot x(t) = (p^2 + \omega_K^2) \sin \omega_K t$$

The transfer function of system  $\Phi(p) = p^2 + \omega_K^2 / \Delta(p)$ , by virtue of condition (5) corresponds to an optimal system, since it contains the  $K(D)$  image of the action  $f(t)$  as a multiplier ( $p^2 + \omega_K^2$  is the  $K(D)$  image of  $f(t) = \sin \omega_K t$ ).

Then, according to condition (4), the function  $|\Phi(j\omega)|^2$  and  $S_f(\omega)$  will respectively have the form of Figure 1.

The product of the function  $|\Phi(j\omega)|^2 S_f(\omega)$  equals zero, since  $|\Phi(j\omega)|^2 \geq 0$  when  $\omega \neq \omega_K$ ,  $|\Phi(j\omega)|^2 = 0$  when  $\omega = \omega_K$

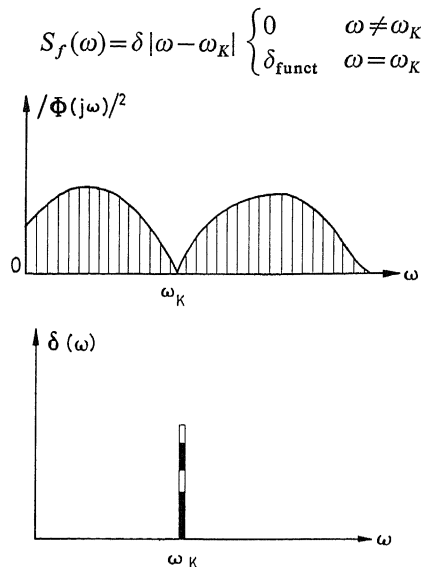


Figure 1

#### Generalization of $K(D)$ Image Theory for the Case of Statistically Given Disturbances $f(t)$

The  $K(D)$  image theory expounded in the works of Kulebakin, was developed for the case of a disturbance  $f(t)$ , preset as a determined function of time  $t$ , particularly for the class of functions which permit approximation of  $f(t)$ , as accurate as desirable, by integrals of linear differential equations, homogeneous and having constant coefficients. Shannon<sup>9</sup> has shown that a very broad class of functions, with the exception of hyper-transcendental functions and  $\xi$  functions, may also be approximated by the solutions of homogeneous differential equations with constant coefficients.

The need to develop statistical methods in the theory of invariance and in particular in the case of  $K(D)$  images is explained by the following. The theory of invariance with an accuracy up to  $\varepsilon$  depends essentially upon the form of  $f(t)$ . The absolute invariance of automatic control systems in the case when the transfer function, as a function of  $f(t)$  equals zero, is generally speaking real for any  $f(t)$ , constrained with respect to the modulus, in particular in relation to those on which information is missing.

In the case of the  $K(D)$  image the effect of absolute invariance may only be observed for a completely defined function  $f(t)$ , knowledge of which, as a determined function of  $t$ , must be available with a probability of 1. Thus essential for the theory of invariance is the knowledge of  $f(t)$ , which is necessary in different cases with a probability from 0 to 1, particularly when investigating invariance with an accuracy up to  $\varepsilon$ . In the case when  $f(t)$  is given in a probabilistic sense, the effect of invariance—particularly from the viewpoint of the  $K(D)$  image theory—was not examined, and the theory of invariance itself is not developed at the present time. An attempt is made below to apply the theory of statistical optimization to the determination of the statistical probabilistic conditions of automatic control systems.

invariance, and generalize the theory of  $K(D)$  images to this case. Henceforward, as previously, we are examining the effect of invariance, the class of statistical actions  $f(t)$  and control systems relating only to stationary systems and stationary actions  $f(t)$ .

*Approximate Conditions of Optimization Using the  $K(D)$  Image in the Case when Dispersion is Present*

In the well-known works of Kolmogorov<sup>10</sup> and others it is shown that any stationary random process may be represented as the limit of a sequence of processes with a discrete spectrum. The general expression of a stationary random process  $f(t)$  in this case may be as follows:

$$f(t) = \sum_{K=1}^n a_K \sin(\omega_K t + \varphi_K) \quad (6)$$

where  $a_1, a_2, a_3, \dots, a_K, \dots, a_n$  are uncorrelated random magnitudes with mean value zero, i.e.,

$$\begin{aligned} M_{a_i} &= 0 & i &= 1, 2, \dots, n \\ M_{a_i} M_{a_j} &= 0 & i &\neq j \end{aligned}$$

where  $M$  is the sign of the mathematical expectation.

It is also known<sup>6, 10</sup> that for each stationary process  $f(t)$  it is possible to indicate a number  $\varepsilon$  as small as desired and an observation time range  $T$  thereof, as large convenient, for as which there exist such pairwise uncorrelated random magnitudes  $a_1, a_2, \dots, a_n$  that the completeness of approximation to the series  $\sum_{K=1}^n a_K \sin(\omega_K t + \varphi_K)$ , determined by the mean-square difference, will be such that

$$M|f(t) - \sum_{K=1}^n a_K \sin(\omega_K t + \varphi_K)|^2 \leq \varepsilon$$

It has thus been shown that each stationary random process  $f(t)$  can be approximated as accurately as desired by the sum of harmonic oscillations with random uncorrelated amplitude and phase. Most essential henceforward is the fact that  $\omega_K$  characterizes the constant frequencies of the process  $f(t)$ .

For the above series the correlation function  $R_f(\tau)$  has, as it is known, the form

$$R_f(\tau) = \sum_{K=1}^n \frac{a_K}{2} \cos \omega_K \tau; \quad (M\{f(t)\} = 0)$$

where  $\omega_1$  is the lower frequency of the spectrum of the random process, equal to  $\omega_1 = 2\pi/\tau_{\max}$ ,  $\tau_{\max}$  is an interval of time, beginning with which  $|R_f(\tau)| < \xi |R_f(0)|$  where  $\xi$  is usually taken to equal 0.05.

For the  $R_f(\tau)$  under consideration, the spectral density  $S_f(\omega)$  represents a discontinuous function, consisting of  $\delta$  functions of the form

$$S_f(\omega) = \sum_{K=1}^n \frac{a_K^2}{2} \delta(\omega - |\omega_K|) \quad (7)$$

By virtue of the foregoing, the condition of an optimal control system is given by the expression

$$|\Phi_{\text{opt}}(j\omega)|^2 S_f(\omega) = 0$$

or on the basis of (7)

$$|\Phi_{\text{opt}}(j\omega)|^2 \cdot \sum_{K=1}^n \frac{a_K}{4} \delta(\omega - |\omega_K|) = 0$$

Since the second co-multiplier of (7) characterizes the spectral density of some periodic function, the expression obtained may be written in the form

$$\begin{aligned} \Phi_{\text{opt}}(p) \cdot a_1 \sin(\omega_1 t + \varphi_1) + \Phi_{\text{opt}}(p) \cdot a_2 \sin(\omega_2 t + \varphi_2) + \dots \\ + \Phi_{\text{opt}}(p) a_n \sin(\omega_n t + \varphi_n) = 0 \end{aligned}$$

In this expression the magnitudes  $a_1, a_2, \dots, a_n$  and  $\varphi_1, \varphi_2, \varphi_n$  are random, undetermined uncorrelated magnitudes,  $\omega_K$  are constants for the given  $f(t)$ . For determination of  $\Phi_{\text{opt}}$  the fact that  $a_K, \varphi_K$  are unknown is not essential, since  $\Phi_{\text{opt}}(p)$ , being the  $K(D)$  image of  $f(t) = \sum_{K=1}^n a_K \sin(\omega_K t + \varphi_K)$  is only determined by the frequency parameter  $\omega_K$ . Since  $\Phi_{\text{opt}}(p)$  for each partial frequency  $\omega_K$  of the spectrum equals  $p^2 + \omega_K^2$ , the following will be the general expression of  $\Phi_{\text{opt}}(p)$

$$\Phi_{\text{opt}}(p) = \left\{ \prod_{K=1}^n (p^2 + \omega_K^2) \right\} \Phi_0(p)$$

where  $\Phi_0(p)$  is the remaining comultiplier of the function  $\Phi_{\text{opt}}(p)$  after the removal from it of  $\prod_{K=1}^n (p^2 + \omega_K^2)$ .

The general problem of the approximate optimization  $\Phi_{\text{opt}}(p)$  of a system in the presence of a random stationary disturbance  $f(t)$  is thus solved with the assistance of the  $K(D)$  image. Expansion of the stationary random process  $f(t)$  into series (6) is a complex problem and it should be carried out on the basis of a preliminary examination of the process  $f(t)$ . So henceforward consideration is given to an assumed case in which the process  $f(t)$  can be characterized by the presence of several main periodic oscillations in the spectrum. In this case the construction of systems satisfying the condition of the  $K(D)$  image is facilitated by the limitation of  $n$ . In a number of practical examples of the use of the  $K(D)$  image for dynamic systems of the damping type, the conditions of the  $K(D)$  image are approximately satisfied only for one  $n = 1$ . The conditions of search of systems satisfying the requirements of  $K(D)$  images may be effected on the basis of the statistical properties of  $f(t)$ . In the above case the automatic control system under consideration must satisfy the condition

$$K(D) \cdot \sum_{K=1}^n a_K \sin(\omega_K t + \varphi_K) = 0 \quad (8)$$

Noting that the  $K(D)$  image is itself invariant to random magnitudes of the series  $f(t)$  with random amplitude  $a_K$  and phase  $\varphi_K$ , and depends only on the determined values of  $\omega_K$ , we shall find the  $K(D)$  image for  $\varphi_K = \pi/2$  and  $a_K = a^2/2$  ( $t = \tau$ ).

Condition (8) will then have the form

$$K(D) \sum_{K=1}^n \frac{a_K}{2} \cos \omega_K \tau = 0$$

or  $K(D) R(\tau) = 0$  where  $R(\tau)$  is the correlation function of  $f(t)$ . Thus the condition of the invariance of the system to the disturbance  $f(t)$ , obtained on the basis of the theory of the  $K(D)$  image, is equivalent to its invariance to the correlation function  $R(\tau)$  of disturbance  $f(t)$ . The conclusion obtained is based on the expression of the stationary random process  $f(t)$  (with a definite degree of accuracy) by a discrete

Kolmogorov series<sup>6</sup>, for which the corresponding spectral density is also the sum of discrete values in the form of  $\delta$  functions. The possibility of using the discrete series (6) determines the applicability of the formula obtained for the case of an  $f(t)$  given by continuous graphs of spectral density.

The condition of invariance to a random function, analogous to the condition derived above, can be obtained if the random function is not expanded into a Kolmogorov series, as was done above, but into a canonical series<sup>5\*</sup>.

The random function  $f(t)$  can be represented by its canonical expansion

$$f(t) = m_f(t) + \sum_v V_v f^0(t)$$

where  $m_f(t)$  is the mathematical expectation of  $f(t)$ , which will henceforward be put equal to zero,  $V_v$  are uncorrelated centred random magnitudes, coefficients of the canonical expansion, and  $f^0(t)$  the coordinate functions of the canonical expansion.

The random coefficients  $V_v$  in the general form of canonical expansion of a random function are determined by the formula<sup>5</sup>

$$V_v = \bar{\Omega}^v F^0(t)$$

where  $\Omega^{(v)}$  are arbitrary linear functionals, which must satisfy the conditions of biorthogonality for the mutual 'non-correlatedness' of the magnitude  $V_v$ ;  $f^0(t)$  is a centred random function.

The condition of invariance of a control system to disturbance  $f(t)$  will be written in the form:

$$\Phi_{\text{opt}}(p) \cdot F^0(t) = 0 \quad (\Phi_{\text{opt}}(p) \neq 0) \quad (9)$$

The coordinate functions  $f(t)$  in the general form of canonical expansion of a random function are determined from the formula

$$f_v^0(t) = \frac{1}{D_v} \Omega_\tau^{(v)} R_f(\tau)$$

where  $D_v$  is the dispersion of an elementary random function, and  $\Omega_\tau^{(v)}$  is an arbitrary linear functional, the lower index of which signifies that this functional is applied to  $R_f(t, \tau)$ , viewed as a function  $\tau$  at a fixed value of  $t$ .

Substituting into (9) the values of the coordinate functions and of coefficients  $V_v$

$$\Phi_{\text{opt}}(p) \cdot \sum_v \bar{\Omega}^v f_v^0(t) \cdot \frac{1}{D_v} \Omega_\tau^{(v)} R_f(\tau) = 0 \quad (10)$$

( $\bar{\Omega}$  is a functional conjugate with  $\Omega$ ). The expression (10) is represented in the form

$$\sum_v \bar{\Omega}^{(v)} f_v(t) \cdot \frac{1}{D_v} \Phi_{\text{opt}}(p) R_f(\tau) = 0 \quad (11)$$

For the identical equality of (11) to zero it is necessary and sufficient with  $f^0(t) \neq 0$ ,  $\Phi_{\text{opt}}(p) \neq 0$ ,  $R_f(\tau) \neq 0$  that  $\Phi_{\text{opt}}(p)$  be the  $K(D)$  image of the correlation function  $R(\tau)$  or contain it as a co-multiplier.

However, it should be noted that the representation of random processes by a spectral series (or canonical expansion) will practically always have a limited number of terms. This constraint causes the appearance of non-zero deflections on the output of the 'invariant' system (non-absolute invariance). The evaluation of this relation has its own significance and is not examined here.

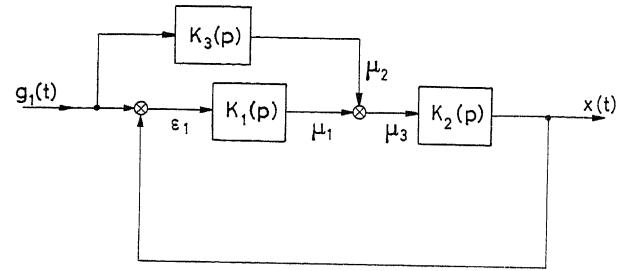
\* The idea of this solution belongs to A. S. Shatalov

### Combined Tracking Systems with Variable Structure

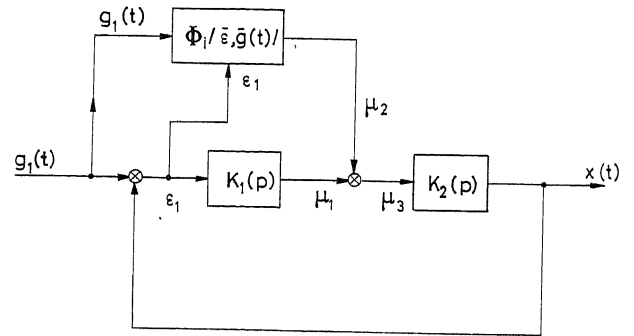
Combined tracking systems are one of the most significant spheres of application of the principle of invariance in automatic control. In the combined system [Figure 2 (a)], reproduction of the control action  $g(t)$  is implemented with the aid of a two-channel system or a system with two cycles: an open-loop cycle  $\mu_2(p) = K_3(p) g(p)$  and a closed-loop cycle

$$x(p) = \frac{K_1(p) K_2(p)}{1 + K_1(p) K_2(p)} g(p)$$

where  $K_1(p)$ ,  $K_2(p)$  are the transfer functions of the elements of the closed-loop cycle,  $K_3(p)$  the transfer function of the open-loop cycle,  $\mu_2$  the output coordinate of the open-loop cycle and  $x$  the controlled coordinate.



(a)



(b)

Figure 2

The transient processes in such systems can be described by the linear non-homogeneous differential equation  $M(p)\varepsilon = N(p)g(t)$  where  $M(p)$  and  $N(p)$  are operator polynomials relative to  $p$ ,  $p \equiv d/dt$ ,  $\varepsilon$  is the error signal. The independence of the error signal from the control action  $g(t)$  is usually determined by the condition

$$N(p) = 0 \quad (12)$$

in this case the forced component  $\varepsilon_{\text{force}}(t)$  of the general solution of the equation of the system is conformally equal to zero. The links with respect to the controlling action  $g(t)$  are selected in such a way as to satisfy condition (12). This is usually achieved by making the coefficients of the polynomial  $N(p)$  consist of the difference of two magnitudes, one of which is determined by the disturbance effect (parameters of the open-loop cycle). It is practically impossible to satisfy condition (12) accurately.



An attempt will be made to solve this problem in another way. A tracking system will be constructed in such a way that the  $n$ -dimensional phase plane of a normal system of non-homogeneous differential equations, by which it is described relative to  $\varepsilon$ , where  $\bar{\varepsilon} = (\varepsilon_1, \varepsilon_2, \dots, \varepsilon_n)$ , with control effect  $g(t)$  of an arbitrary type, contains some  $(n-1)$ -dimensional hyperplane  $S$ , and it will be required that the motion of the state point in  $S$  be described by a system of homogeneous differential equations. Then, if the state point under any initial conditions and for any forms of  $g(t)$  terminates its motion in this  $(n-1)$ -dimensional diversity of  $S$ , the error of signal  $\varepsilon$  will always tend to zero ( $\varepsilon \rightarrow 0$ ) for any  $g(t)$ . In other words, the controlled coordinate  $x(t)$  will reproduce any continuous  $g(t)$  without static error, and the requirements for the operator  $N(p)$ , determined by condition (12), will be absent. If the function  $g(t)$  has a discontinuity at some moments, then slight dynamic errors will appear at these moments. An attempt will be made to solve this problem, using the principles of construction of variable-structure automatic control systems<sup>12</sup>.

#### Conditions of Invariance in Combined Tracking Systems with Variable Structure

In the domain,  $G$ , of an  $n$ -dimensional space  $\varepsilon_1, \dots, \varepsilon_n$  let the motion of a dynamic system be described by a system of non-homogeneous differential equations with a discontinuous right-hand side

$$\frac{d\bar{\varepsilon}}{dt} = \bar{f}(\bar{\varepsilon}, \bar{g}(t)) \quad (13)$$

Here

$$\bar{\varepsilon} = (\varepsilon_1, \dots, \varepsilon_n), \bar{g} = (g_1, \dots, g_m), \bar{f} = (f_1, \dots, f_n)$$

$$f_i = \varepsilon_{i+1} \quad (i = 1, 2, \dots, n-1)$$

$$f_n = - \sum_{i=1}^n a_i \varepsilon_i + \sum_{i=1}^m \psi_i(\bar{\varepsilon}, \bar{g}(t)) g_i(t)$$

where

$$\psi_i(\bar{\varepsilon}, \bar{g}(t)) = \begin{cases} b_i \text{ for } \left( \sum_{j=1}^n c_j \varepsilon_j \right) g_i(t) > 0^\dagger, \\ b_i^* \text{ for } \left( \sum_{j=1}^n c_j \varepsilon_j \right) g_i(t) < 0 \end{cases} \quad i = 1, 2, \dots, n$$

$a_i, b_i, b_i^*, c_j$  are constants,  $g_i(t)$  is a function defined and continuous on the whole time interval  $t$ . Let the hyperplane  $S$ , set by the equation  $\sum_{j=1}^n c_j \varepsilon_j = 0$  divide the domain  $G$  into sub-domains  $G^+ \left( \sum_{j=1}^n c_j \varepsilon_j > 0 \right)$  and  $G^- \left( \sum_{j=1}^n c_j \varepsilon_j < 0 \right)$ , in which the vector function  $\bar{f}(\bar{\varepsilon}, \bar{g}(t))$  of system (13) is constrained and for any constant value of time  $t$  on the approach to  $S$  from  $G^+$  and  $G^-$  there exist its limit values  $\bar{f}^+(\bar{\varepsilon}, \bar{g}(t))$  and  $\bar{f}^-(\bar{\varepsilon}, \bar{g}(t))$ . On the approach of the solution  $\bar{\varepsilon}(t)$  to some domain  $U \subset S$  let the vector functions  $\bar{f}^+$  and  $\bar{f}^-$  be directed towards the hyper-

plane  $S$  ( $f_N^+ > 0, f_N^- < 0$ , where  $f_N^+$  and  $f_N^-$  are the projections of the vectors  $\bar{f}^+$  and  $\bar{f}^-$  on to the normal to the hyperplane  $S$ , directed from  $G^-$  to  $G^+$ ). Then, when  $\bar{\varepsilon}(t)$  hits  $U$  there arises the so-called sliding mode and the solution of system (13) does not depend on  $a_i, b_i, b_i^*, g_i(t)$ . In fact in this case, as shown by Filippov<sup>13</sup>, in the domain  $U^-$  there exists a solution  $\bar{\varepsilon}(t)$  of system (13), and the vector  $d\bar{\varepsilon}/dt = \bar{f}^0(\bar{\varepsilon}, \bar{g}(t))$ , where  $\bar{f}^0 = (f_1^0, \dots, f_n^0)$ , lies in the hyperplane  $S$  and is determined by the values of the vector functions  $\bar{f}^+$  and  $\bar{f}^-$ .

From the condition that  $\bar{f}^0(\bar{\varepsilon}, \bar{g}(t)) \in S$  there follows the linear relationship of the components of the vector  $\bar{f}^0$

$$\sum_{j=1}^n c_j f_j^0 = 0 \quad (14)$$

where  $f_j^0$  is the  $j$ th component of the vector  $\bar{f}^0$  whence

$$f_n^0 = - \frac{1}{c_n} \sum_{j=1}^{n-1} c_j f_j^0 \quad (15)$$

Hence the solution of system (13) for  $\bar{\varepsilon}(t) \in U$  coincides with the solution of the linear system of homogeneous differential equations

$$\frac{d\bar{\varepsilon}}{dt} = \bar{f}^0(\bar{\varepsilon}) \quad (16)$$

Here

$$\bar{\varepsilon} = (\varepsilon_1, \dots, \varepsilon_n)$$

$$f_j = \varepsilon_{j+1} \quad (j = 1, 2, \dots, n-1), f_n^0 = - \frac{1}{c_n} \sum_{j=1}^{n-1} c_j \varepsilon_{j+1}$$

$c_j$  are constants.

Obviously the solution of system (16) does not depend on  $a_i, b_i, b_i^*, g_i(t)$ . Use will be made of this property of the solution of the system of non-homogeneous differential equations with a discontinuous right-hand side for the construction of a combined tracking system with variable structure.

Let the structure, selected in a definite way, of the open-loop cycle of a combined tracking system [Figure 2(b)] change stepwise on some hyperplane  $S = \sum_{j=1}^n c_j \varepsilon_j = 0$  in such a way that the movement of this servosystem is described by a system of non-homogeneous differential equations with a discontinuous right-hand side (13), where  $\psi_i(\bar{\varepsilon}, \bar{g}(t)) = F[\Phi_i(\bar{\varepsilon}, \bar{g}(t))]$

$$\Phi_i(\bar{\varepsilon}, \bar{g}(t)) = \begin{cases} K_i \text{ for } \left( \sum_{j=1}^n c_j \varepsilon_j \right) g_i(t) > 0^\dagger \\ K_i^* \text{ for } \left( \sum_{j=1}^n c_j \varepsilon_j \right) g_i(t) < 0 \end{cases} \quad (i = 1, 2, \dots, n)$$

$K_i, K_i^*$  are constants, determined by the open-loop cycle parameters. It is assumed (a) that the domain  $U$  exists, it includes the origin of the coordinates, and the solution of the system of

<sup>†</sup> In the case  $\left( \sum_{j=1}^n c_j \varepsilon_j \right) g_i(t) = 0$

$$\psi_i(\bar{\varepsilon}, \bar{g}(t)) = b_i \text{ for } \left( \sum_{j=1}^n c_j \varepsilon_j \right) g_i(t) \rightarrow +0,$$

$$\psi_i(\bar{\varepsilon}, \bar{g}(t)) = b_i^* \text{ for } \left( \sum_{j=1}^n c_j \varepsilon_j \right) g_i(t) \rightarrow -0$$

<sup>†</sup> In the case  $\left( \sum_{j=1}^n c_j \varepsilon_j \right) g_i(t) = 0$

$$\Phi_i(\bar{\varepsilon}, \bar{g}(t)) = K_i \text{ for } \left( \sum_{j=1}^n c_j \varepsilon_j \right) g_i(t) \rightarrow +0$$

$$\Phi_i(\bar{\varepsilon}, \bar{g}(t)) = K_i^* \text{ for } \left( \sum_{j=1}^n c_j \varepsilon_j \right) g_i(t) \rightarrow -0$$

differential equations (16) satisfies the given requirements on the quality of the process of control (control time and maximum dynamic error of the system must not exceed certain predetermined values); (b) there exists a sufficiently large domain of initial conditions under which the solution of the system of equations (13) hits the domain  $U$ ; (c) in the domain  $U$  there do not exist trajectories serving as sectors of limit cycles with a partially sliding regime.

Then the solution of the initial non-homogeneous system of differential equations (13) will depend on the controlling action  $g_1(t)$  and the parameters of the closed-loop and open-loop cycles only up to the moment when  $\bar{\varepsilon}(t)$  hits the domain  $U$ , where the solution coincides with the solution of the linear system of homogeneous differential equations (16) and depends only on the coefficients  $c_j$ . Thus in this case, in the reproduction of the controlling actions  $g_1(t)$  the magnitude  $\varepsilon_1 \rightarrow 0$  on a finite interval of time  $t$ , and the controlled coordinate  $x(t)$  reproduces  $g_1(t)$  without static error. The quality of the process of control in such systems depends loosely on the variation of the parameters of the open-loop and closed-loop cycles, since the solution  $\varepsilon(t)$  depends on these parameters only until it hits the domain  $U$ . It must be noted that in the systems under examination, the open-loop cycle for  $g_1(t) \neq 0$  in some cases exerts an influence on the stability of the tracking system. In particular, as an example will demonstrate, even when the change of the parameters of the closed-loop cycle leads to the loss of the stability of the closed loop, then on the whole for  $g_1(t) \neq 0$  the open-loop cycle with variable structure will ensure, in some domain of initial conditions, the stable operation of the tracking system. The above-listed properties of combined tracking systems with variable structure advantageously distinguish them from ordinary linear combined tracking systems.

### An Example of a Combined Tracking System with Variable Structure

Let the equations of the individual components of a combined servosystem with variable structure have the form

$$\mu_1 = k_1 \varepsilon_1, T_1 T_2 \ddot{g}_1(t) + (T_1 + T_2) \dot{g}_1(t) + g_1(t) = k_2 \mu_3$$

$$\mu_2 = \Phi_1(\bar{\varepsilon}, g, (t)) g_1(t)$$

$$\Phi_1(\bar{\varepsilon}, g_1(t)) = \begin{cases} +K \text{ for } (c_1\varepsilon_1 + c_2\varepsilon_2) g_1(t) < 0^* \\ -K \text{ for } (c_1\varepsilon_1 + c_2\varepsilon_2) g_1(t) > 0 \end{cases}$$

$$g_1(t) - x(t) = \varepsilon_1$$

where  $k_1, k_2, K, T_1, T_2, c_1, c_2$  are constants. The block diagram of the system is depicted in *Figure 3(a)* and *(b)*. In this case the combined tracking system, after the elimination of the intermediate coordinates  $\mu_1, \mu_2, \mu_3, x$  is described by the following system of non-homogeneous differential equations with a discontinuous right-hand side:

$$\frac{d\bar{\varepsilon}}{dt} = \bar{f}(\bar{\varepsilon}, \bar{g}(t))$$

$$\text{Here } \bar{\varepsilon}=(\varepsilon_1, \varepsilon_2), \bar{g}=(g_1, g_2, g_3), f=(f_1, f_2) \quad (17)$$

$$f_1 = \varepsilon_2, f_2 = -2b\bar{\varepsilon}_2 - \omega_0^2 \varepsilon_1 + g_3(t) + 2bg_2(t) + \psi_1(\bar{\varepsilon}, g_1(t))g_1(t)$$

where

$$2b = \frac{T_1 + T_2}{T_1 T_2}, \quad \omega_0^2 = \frac{1 + k_1 k_2}{T_1 T_2},$$

$$\psi_1(\varepsilon, g_1(t)) = \frac{1 + \Phi_1(\bar{\varepsilon}, g_1(t))}{T_1 T_2},$$

$$g_2(t) = \frac{dg_1(t)}{dt} \quad g_3(t) = \frac{dg_2(t)}{dt}$$

We shall examine the behaviour of a combined tracking system with variable open-loop structure which reproduces various controlling actions  $g_1(t)$ , while the parameters of the transfer function of the closed-loop cycle  $K_2(p)$  can be chosen within wide limits.

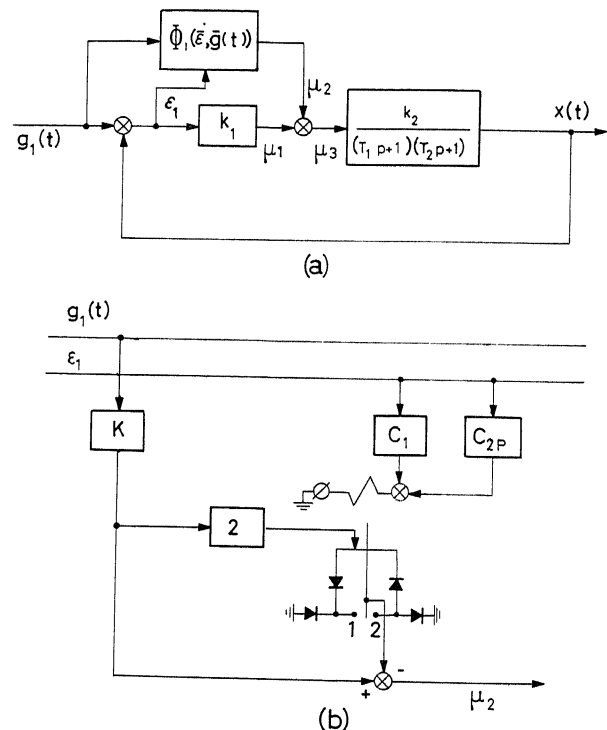


Figure 3

The phase-plane method is used for analysis of the system. Let the controlling action  $g_1(t) = A$ , where  $A$  is a constant and the parameters of the tracking system  $k_1, k_2, T_1, T_2, K$  are selected in such a way as to satisfy the following conditions:

$$K \cdot k_2 > 1 \quad (18)$$

$$b^2 > \omega_0^2 \quad (19)$$

$$\frac{c_1}{c_2} = -b - \sqrt{(b^2 - \omega_0^2)^2} \quad (20)$$

Then for  $g_1(t) > 0$  the phase plane of the system will have the form shown in *Figure 4(a), (b) and (c)*. In this case, under any initial conditions the state point will tend to hit the straight

\* For  $(c_1 \varepsilon_1 + c_2 \varepsilon_2) g_1(t) = 0$

$$\Phi_i(\bar{\varepsilon}, g_1(t)) = +K \text{ for } (c_1 \varepsilon_1 + c_2 \varepsilon_2) g_1(t) \rightarrow +0$$
$$\Phi_i(\bar{\varepsilon}, g_1(t)) = -K \text{ for } (c_1 \varepsilon_1 + c_2 \varepsilon_2) g_1(t) \rightarrow -0$$

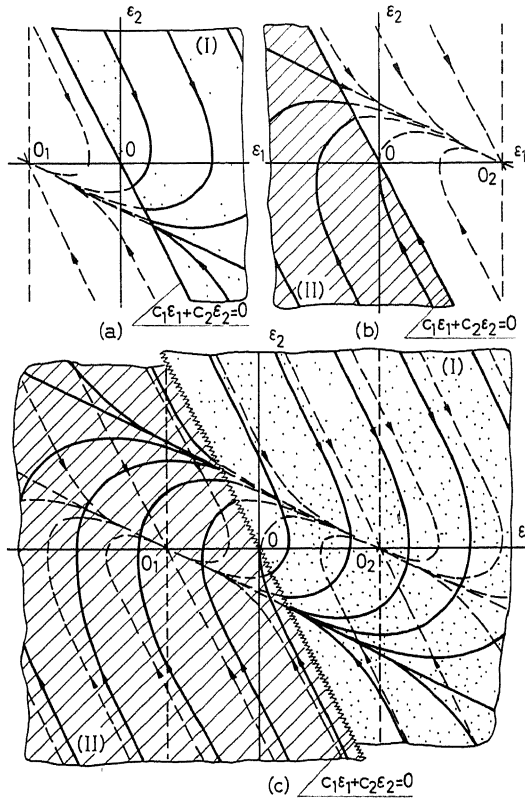


Figure 4

line  $c_1 \varepsilon_1 + c_2 \varepsilon_2 = 0$  which serves as the boundary of discontinuity of the right-hand side of eqn (17) while on the boundary of discontinuity the vector functions  $\bar{f}^+$  (sheet I) and  $\bar{f}^-$  (sheet II) are always directed towards this straight line and, hence, when the state point hits it the solution of eqn (17) coincides with the solution of the linear homogeneous differential equation

$$\frac{d\bar{\varepsilon}}{dt} = f^0(\bar{\varepsilon}) \quad (21)$$

Here

$$\bar{\varepsilon} = (\varepsilon_1, \varepsilon_2), \bar{f}^0 = (f_1^0, f_2^0)$$

$$f_1^0 = \varepsilon_2, f_2^0 = \frac{c_1}{c_2} \varepsilon_2$$

Thus the right-hand side of the equation determines the motion of the system only up to the moment when the state point hits the boundary of discontinuity, and then the motion of the system can be reflected by an equation without a right-hand side (21) or, after the appropriate transforms, by the equation

$$c_1 \varepsilon_1 + c_2 \varepsilon_2 = 0 \quad (22)$$

In this case, therefore, static error will be absent. We shall follow the variation of the static and dynamic properties of the system as the parameters of the transfer function of the closed-loop cycle  $K_2(p)$  vary.

Let the parameters of the closed-loop cycle vary in such a way that the closed loop of the system becomes unstable, e.g., consider that the sign is altered in front of the term  $2b\varepsilon_2$  in eqn (17).

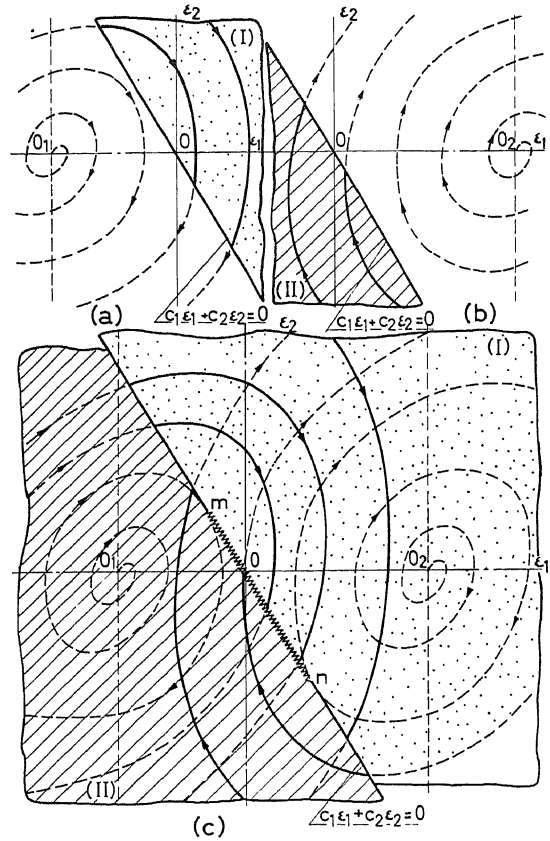


Figure 5

In this case the system will also become unstable for any parameters of the linear transfer function of the open-loop cycle. When there is an open-loop cycle with a variable structure, the phase plane will have the form shown in Figure 5. As before, the state point, under any initial conditions, will hit the straight line (22), on which there exists a finite length  $mn$  which includes the origin of the coordinates 0, where the conditions of the existence of a sliding mode are satisfied. The tracking system will thus be stable. For a particular set of initial conditions the process will run without overshoot, and as before there will be no static error. Thus the variable-structure tracking systems under consideration are insensitive in relation to variation of the system parameters.

It is not difficult to show that for  $g(t) < 0$  all the examined properties of the combined tracking system with variable structure will remain unchanged. We shall consider whether these properties of the system are preserved when reproducing other forms of controlling actions, e.g.,  $g_1(t) = \alpha t$ ,  $Ae^{\alpha_1 t}$  where  $\alpha, \alpha_1$  are constants.

In this case one will be dealing with a non-stationary phase plane. By examining the field of the tangents to the phase trajectories for various fixed moments of time  $t$ , the change of the directions of the vector functions  $f t$  and  $f$  can be followed and thus the answer given to the question of the existence of a section of sliding mode  $mn$  on the straight line (22) and the landing of the state point on this section.

Let the control action  $g_1(t) = \alpha t$ .

We shall examine the static and dynamic properties of a tracking system for the first case of combination of closed-loop

cycle parameters. For the instant  $t = 0$  [Figure 6(a)] the direction of the vector functions  $\vec{f}^+$  and  $\vec{f}^-$  in the vicinity of the straight line (22) is such that the section of sliding mode  $mn$  on straight line (22) is everywhere absent. However, with the time, beginning with some  $t = t_1$ , the field of the tangents to the phase trajectories changes in such a way that the vector functions  $\vec{f}^+$  and  $\vec{f}^-$  in the vicinity of the straight line (22) are everywhere directed towards this straight line [Figure 6(b)]. Since, with time the inclination of the tangents to the phase trajectories is deformed in such a way that at the limit it tends towards straight lines [Figure 6(c)], the above-mentioned static and dynamic properties will also be preserved when the system is reproducing a controlling action of the kind under review. Let the controlling effect be  $g_1(t) = Ae^{\alpha t}$ . From the analysis of the variation of the fields of the tangents for various instants  $t$ , it follows that even when reproducing a transcendental control action, static error is absent.

With the aid of an electronic simulator we shall study the behaviour of such a combined tracking system with a variable structure in the reproduction of controlling actions of the form

$$g_1(t) = A, \alpha t + A_1, A_2 e^{\alpha_1 t} \text{ where } A, \alpha, A_1, \alpha_1, A_2$$

are constants.

Let the parameters of the tracking system equal

$$T_1 = 1, T_2 = 1, k_1 = 1, k_2 = 1, K = 2$$

As follows from the oscillograms in Figure 8(a), (b) and (c)

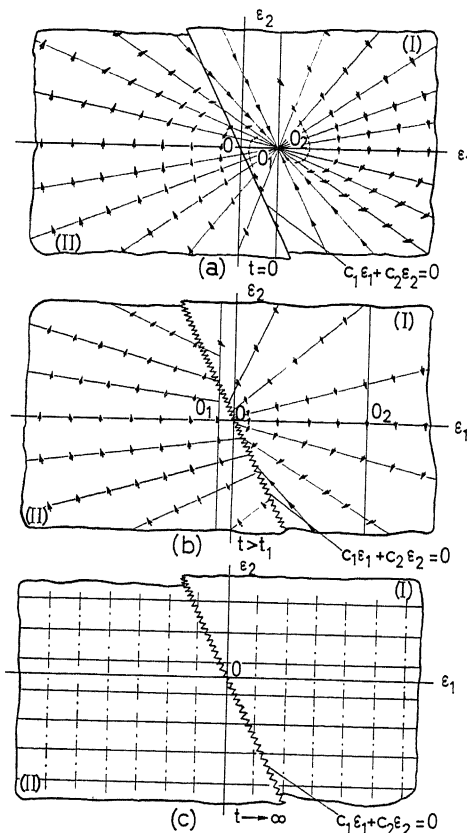


Figure 6

all the controlling actions under review are reproduced without static errors with a good quality of the transient processes.

We shall change any one of the parameters of the controlled plant, e.g.,  $k_2$  from the value  $k_2 = 1$  to  $k_2 = 10$ , and follow the change of the static and dynamic errors of a combined tracking system with variable structure. As can be seen from the oscillograms in Figure 8(d), (e) and (f) the static properties of the system have been preserved in this; as before, it reproduces, without static errors, all the types of control actions under consideration, while the dynamic properties have not suffered any qualitative changes—the time of the transient processes has been slightly reduced.

## Conclusions

The paper considers the invariance of automatic control systems in the presence of statistically given disturbances. The invariance conditions, obtained on the basis of the  $K(D)$  image theory, have been generalized for the case of statistically given disturbances. For stationary systems of automatic control and stationary disturbances  $f(t)$  the conditions of the  $K(D)$  images in relation to the disturbance prove to be equivalent to the condition of the  $K(D)$  image in relation to its correlation function.

A new principle has been proposed for the design of invariant tracking systems in relation to continuous functions of the controlling action, which ensure the absence of static error. It is shown when using an open-loop cycle with variable structure that it is possible to reproduce, without static errors, an extensive

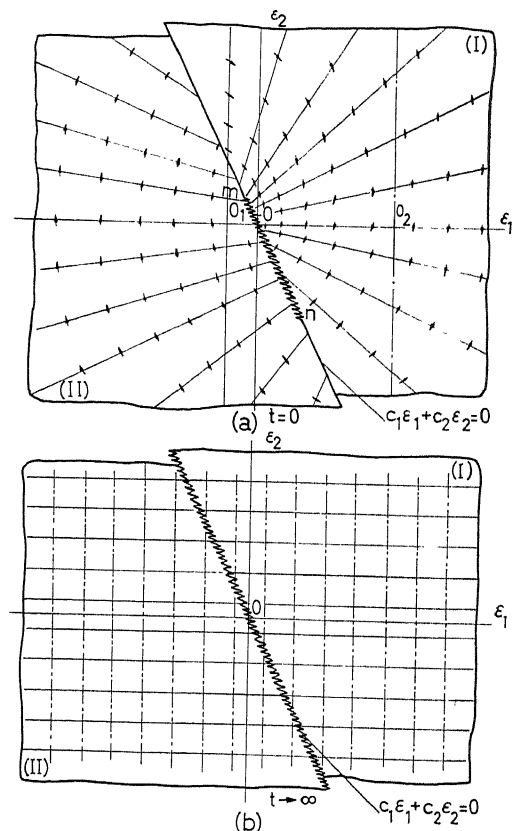


Figure 7

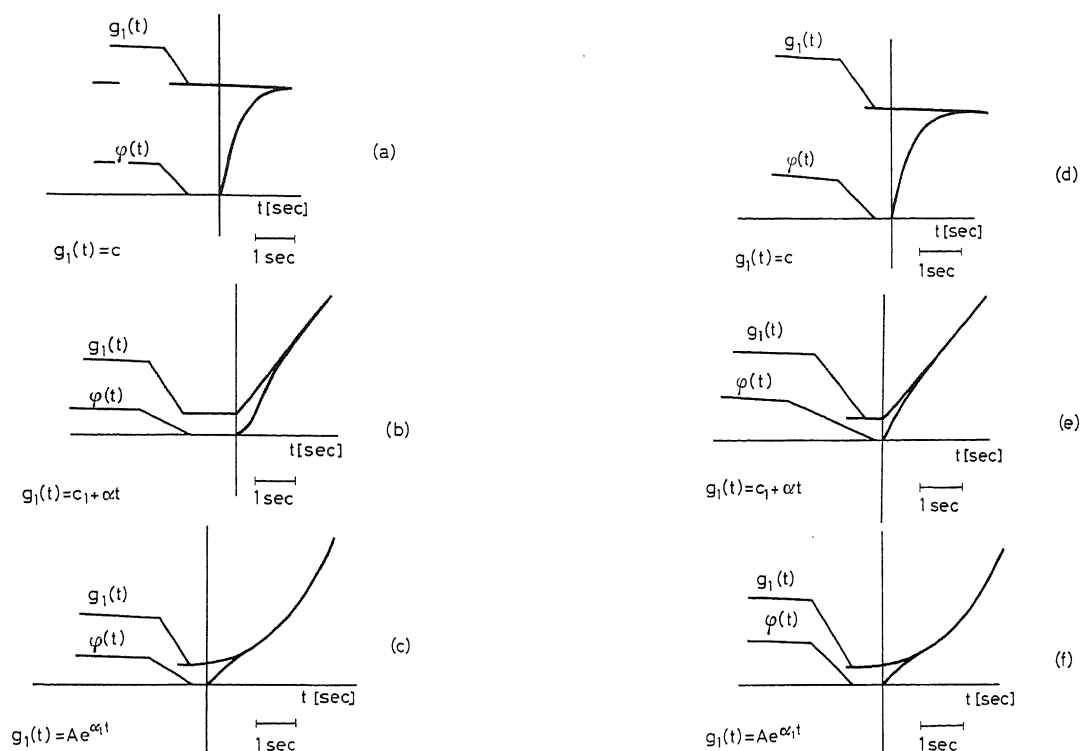


Figure 8

class of controlling-action functions. When selecting the open-loop cycle transfer-action function there is no need to satisfy the classical conditions of invariance, which require the right-hand side of the non-homogeneous differential equation to vanish. This property of the systems under consideration makes it possible to build invariant tracking systems without differentiation of controlling action. The variable-structure combined tracking systems considered are insensitive to the variation of the system parameters within a certain range.

## References

- <sup>1</sup> KULEBAKIN, V. S. *Uspekhi. Mat. Nauk.* 6, No. 5 (1951), 211; *Dokl. Akad. Nauk.* 68, No. 5 (1949); *Dokl. Akad. Nauk.* 77, No. 2 (1951)
- <sup>2</sup> PETROV, B. N., and ULANOV, G. M. Some problems of the theory of combined automatic control systems. *Trud. Sessii Akad. Nauk S.S.S.R.*, No. 5 (1956)
- <sup>3</sup> PETROV, B. N. The principle of invariance and the condition of its use in designing linear and non-linear systems. *Automatic and Remote Control. Proc.* 1. 1960. London; Butterworths
- <sup>4</sup> WIENER, N. *Extrapolation, Interpolation and Smoothing of Stationary Time Series.* 1949. New York
- <sup>5</sup> PUGACHEV, V. S. *The Theory of Random Functions.* 1957. Moscow; State Publishing House of Physico-Mathematical Literature
- <sup>6</sup> SOLODOVNIKOV, V. V. *The Statistical Dynamics of Linear Automatic Control Systems.* 1960. Moscow; State Publishing House of Physico-Mathematical Literature
- <sup>7</sup> IVAKHNENKO, A. G. *Electro-automation. Izd. Akad. Nauk Ukr. S.S.R. Pt. 1, II* (1957)
- <sup>8</sup> ULANOV, G. M. *Disturbance Control.* 1960. Moscow; Gosenergoizdat
- <sup>9</sup> SHANNON, R. *J. Math. Phys. U.S.A.* 20 (1941)
- <sup>10</sup> KOLMOGOROV, A. N. Stationary sequences in a Gilbert space. *MGU.* 2, No. 6 (1941)
- <sup>11</sup> KAZAKOV, I. B. Approximate probability analysis of the accuracy of operation of essentially non-linear systems. *Avtomatika i telemekhanika* XVII, No. 5 (1956)
- <sup>12</sup> EMEL'YANOV, S. V. The use of 'key' type non-linear correcting devices to improve the quality of second-order automatic control systems. *Avtomatika i telemekhanika* XX, No. 7 (1959)
- <sup>13</sup> FILIPPOV, A. F. Differential equations with a discontinuous right-hand side. *Matem. sbornik* No. 1 (1960)

## DISCUSSION

J. E. GIBSON, 168, Drury Lane, West Lafayette, Indiana, U.S.A.

It appears that it would be difficult or impossible to design a system which satisfies the law of eqn (12) of the paper.

S. EMEL'YANOV, in reply

It is, in fact, impossible that the condition (12) be strictly satisfied in the real linear systems, because even very small variations of parameters will lead to its deterioration.

The main point of the paper is a development of such systems which would in some way have properties of linear systems under a strict satisfaction of condition (12) (astatic realization of any analogue

function of control action). At the same time, they should not have the shortcomings of linear systems [difficulties of strict satisfaction of condition (12)]. This problem can be solved by the application of control systems with a variable structure. They can provide the conditions under which the phase space of the system developed will include some  $(n - 1)$  dimensional subspace, where the motion would be expressed by homogeneous differential equations.

In this case the control actions will be realized without statistical errors. Besides, the variations of system parameters in a wide range will not give a stable error and will not deteriorate the transfer function of the process as it follows from the example given in the paper.

# Synthesis of Systems with Fixed Structures of Equivalent Self-adjusting Systems

M. V. MEEROV

## Summary

The paper describes the methods of designing automatic control systems with constant structures and constant parameters, the properties of which are equivalent to one of the adaptive systems.

These systems are based on the structures which admit an unlimited increase of gains without disturbing the stability.

It is shown that the presence of the disturbance at the input of the compensating network (when the disturbance can be measured) or the introduction of the model of the plant to the system provides, for the chosen structure of the whole system, the removal of the influence of external disturbances.

It is concluded that the sensitivity, according to Bode, for the mentioned structures does not depend on the parameters of the plant, nor does its value tend to zero under the corresponding increase of the gain.

## Sommaire

On décrit les méthodes de calcul de systèmes de commande automatique à structures et à paramètres constants, dont les propriétés sont équivalentes à celles de systèmes adaptatifs. Dans ces systèmes, une augmentation illimitée de gain ne perturbe pas la stabilité.

En cas de perturbations mesurables à l'entrée du réseau compensateur, ou de l'introduction du modèle de l'installation commandée leur influence se trouve éliminée avec une structure définie d'avance pour le système global.

Dans ces systèmes, la sensibilité au sens de Bode ne dépend pas des paramètres de l'installation, et sa valeur ne tend pas vers zéro quand le gain augmente d'une manière correspondante.

## Zusammenfassung

Der Aufsatz beschreibt Entwurfsmethoden für automatische Regelsysteme mit fester Struktur und konstanten Parametern, deren Eigenschaften einem der selbststellenden Systeme entsprechen.

Derartige Systeme beruhen auf Strukturen, die eine unbegrenzte Zunahme der Verstärkung zulassen, ohne instabil zu werden.

Wie die Arbeit zeigt, lassen sich die äußeren Störeinflüsse einmal ausschalten, indem man sie als zusätzliches Eingangssignal (wenn die Störung meßbar ist) dem kompensierenden Netzwerk zuführt; zum andern gelingt die Unterdrückung der Störeinflüsse durch Hinzufügung eines Modelles der Strecke zum gewählten Gesamtsystem.

Es wird daraus geschlossen, daß für die erwähnten Strukturen die Empfindlichkeit nach Bode weder von den Parametern der Strecke abhängt noch deren Wert gegen Null strebt, wenn eine entsprechende Zunahme der Verstärkung vorliegt.

The problem of 'self-adjustment' in a control system arises in connection with the fact that in the operating process the characteristics of the controlled plant may vary within a wide range. Under these conditions, the adjustment of the control system, or even its structure, which was entirely expedient for the initial form of the plant's characteristics, may prove to be completely unsatisfactory for the altered characteristics.

A change in the characteristics of the plant to be controlled may be conditioned, basically, by the two most prevalent factors. In the first place, a change in the characteristics may be brought about by external disturbances that are applied to the plant; and in the second place, a change in the plant's characteristics may take place in the course of its operation.

The problem of 'self-adjustment' also comes up in a number of cases where the information regarding the characteristics and properties of the plant is insufficient; it is only known that the plant's characteristics have an extremum for some quality criterion, and the control system's problem consists in a search for this extremum and in maintaining the plant's operating conditions at this extremum. A rather large number of studies (for example, by Feldbaum<sup>1, 2, 4</sup>, and by Doganovskii and Feldbaum<sup>3</sup>) have been devoted to methods of searching for, and adjustment of the system to, the disclosed extremum. In the present paper, methods are considered for the purpose of constructing systems with fixed structures that would maintain the most favourable adjustment, independently of external disturbances, and the character of which may be practically arbitrary. The sole limitation is the one regarding disturbances in accordance with the modulus (absolute value). In the present study, no consideration is given to the method of searching for the extremal condition for some quality criterion. If, however, the extremum is established by some method or other, then the structures examined below maintain this extremum automatically, without the need for a duplicate search.

## Methods of Constructing Control Systems for the Case where the External Disturbances may be Measured

Consider the automatic control system whose block diagram is shown in Figure 1. In this diagram the designation  $w_2(p)$  is given to the transfer function of the controlled plant,  $kw_1(p)$  and  $w_3(p)$  to the transfer functions of the control system and of the stabilizing device, and  $F$  to the external disturbance.  $kw_1$  and  $w_3(p)$  have been selected in such a way

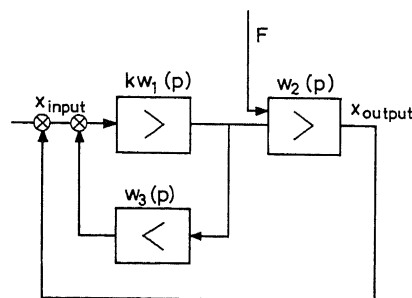


Figure 1

that, in the absence of disturbances  $F$ , the process, which is the most favourable from the point of view of the selected quality criterion, is attained in the case of a sufficiently large gain,  $k$ . Thus, for example, the optimum operating conditions are realized where there is an unlimited increase in the gain, the plant is non-linear, and there is a non-linear feedback in the optimum control circuit with an automatic potentiometer<sup>5, 6</sup>. It is natural for the designed circuit to remain stable when there is an unlimited increase in the gain. The following proof is given: the structure, which is shown in Figure 1, giving no consideration to external disturbances and where  $k$  tends to infinity, is equivalent to the system in Figure 2, where consideration is given to disturbances and where  $k$  tends to infinity. In other words, in order to eliminate the effect of external disturbances that are capable of being measured, they should be supplied to the input of the stabilizing device in the form of an additional signal.

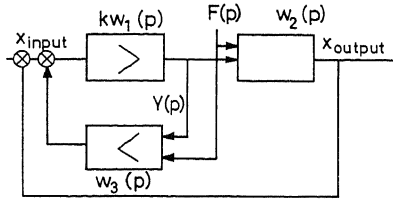


Figure 2

Actually, the transfer function of the system in Figure 1, without calculating the external disturbances, will have the form:

$$k(p) = \frac{x_{\text{output}}(p)}{x_{\text{input}}(p)} = \frac{kw_1(p)}{1 + kw_1(p)w_3(p)} w_2(p) = \frac{kw_1(p)w_2(p)}{1 + kw_1(p)w_3(p) + kw_1(p)w_2(p)} \quad (1)$$

Assume that  $k$  tends to infinity; then,

$$k_{\text{adjusted}}(p) = \frac{w_1(p)w_2(p)}{w_1(p)w_3(p) + w_1(p)w_2(p)} = \frac{w_2(p)}{w_3(p) + w_2(p)} \quad (2)$$

Now, the transfer function for the circuit in Figure 2 is found; one has:

$$Y(p) = kw_1(p) \{x_{\text{input}} - x_{\text{output}} - [Y(p) + F(p)]w_3(p)\} \quad (3)$$

$$x_{\text{output}}(p) = w_2(p) [Y(p) + F(p)] \quad (4)$$

From (3),

$$Y(p) = \frac{kw_1(p)x_{\text{input}}(p) - kw_1(p)x_{\text{output}}(p) - kw_1(p)w_3(p)F(p)}{1 + kw_1(p)w_3(p)} \quad (5)$$

Substituting the value  $Y(p)$  from (5) into (4), one obtains:

$$\begin{aligned} & [1 + kw_1(p)w_3(p)]x_{\text{output}}(p) \\ &= kw_1(p)w_2(p)x_{\text{input}}(p) - kw_1(p)w_2(p)x_{\text{output}}(p) + F(p) \\ & - kw_1(p)w_2(p)w_3(p)F(p) + w_2(p)F(p) + kw_1(p)w_2(p)w_3(p) \end{aligned}$$

from which:

$$x_{\text{output}}(p) = \frac{kw_1(p)w_2(p)x_{\text{input}}(p) + w_2(p)F(p)}{1 + kw_1(p)w_2(p) + kw_1(p)w_3(p)} \quad (6)$$

When  $k$  tends to infinity, one obtains:

$$x_{\text{output}}(p) = \frac{w_1(p)w_2(p)x_{\text{input}}(p)}{w_1(p)w_2(p) + w_1(p)w_3(p)}$$

or

$$\lim_{k \rightarrow \infty} \frac{x_{\text{output}}(p)}{x_{\text{input}}(p)} = \frac{w_2(p)}{w_2(p) + w_3(p)} \quad (7)$$

i.e. exactly the same expression as eqn (2). From what has been obtained it follows that the system in Figure 2, when there is a sufficiently large gain, will behave as a self-adjusting one, in the sense that its characteristics will remain unchanged despite the presence of external disturbances whose character is practically unlimited.

### Methods of Constructing Structures for the Case where it Does Not Appear Possible to Measure Disturbances Directly

Now consider the case where the plant's characteristics change due to the effect of external disturbances, but where it does not appear possible to measure these disturbances. Such a situation is, for all practical cases, highly prevalent. A series of disturbances is difficult to measure, in the first place, because of the properties of the disturbances themselves, and in the second place because of the absence of sufficiently high-speed measuring devices for the measurement of the external disturbances.

The solution of the problem in the given case is carried out in the following fashion. Assume that the plant's characteristics are known for the case where disturbances are absent. For this case, a control system is constructed in such a way that the optimum operating conditions should be attained when there is an unlimited increase in the gain,  $k$ . Strictly speaking, in the absence of disturbances, the system has the form shown by Figure 1. As was indicated earlier in this paper, in the case where  $k$  tends to infinity, one has:

$$k_{\text{adj.}}(p) = [w_2(p)]/[w_3(p) + w_2(p)]$$

Now Figure 3 is plotted. The output of the controlling part of the circuit, which is designated in Figure 3 by the letter  $Y$ , is fed to the input of the real plant and to the input of the model with a transfer function  $w'_2(p)$ . In future,  $w'_2(p)$  is called the transfer function of an ideal plant.

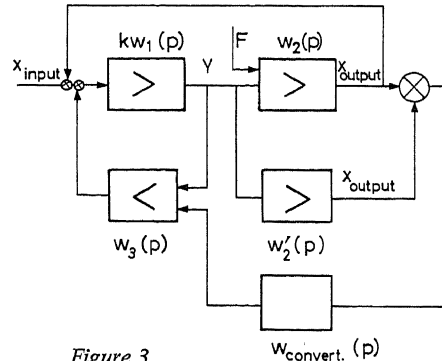


Figure 3

The difference between the outputs of the ideal and real plants is fed through a converting device with a transfer function  $w_{\text{convert}}(p)$ , to the input of the stabilizing device. Now the transfer function of the system in *Figure 3* is found.

$$Y(p) = kw_1(p) \{x_{\text{input}}(p) - x_{\text{output}}(p) - w_3(p) [Y(p) + (x_{\text{output}}(p) - x'_{\text{output}}(p)) w_{\text{convert}}(p)]\} \quad (8)$$

$$x_{\text{output}}(p) = w_2(p) [Y(p) + F(p)] \quad (9)$$

$$x'_{\text{output}}(p) = w_2(p) Y(p) \quad (10)$$

On the basis of (9) and (10), one may write:

$$x_{\text{output}}(p) - x'_{\text{output}}(p) = w_2(p) [Y(p) + F(p) - Y(p)] = w_2(p) F(p) \quad (11)$$

by calculating (11), eqn (8) is written as:

$$Y(p) = kw_1(p) \{x_{\text{input}}(p) - x_{\text{output}}(p) - w_3(p) Y(p) - w_3(p) w_{\text{convert}}(p) w_2(p) F(p)\} \quad (12)$$

From (12), the expression for  $Y(p)$  is found, namely:

$$\boxed{\text{Eqn (13)}}^*$$

By substituting the expression for  $Y(p)$  from (13) in (9), one obtains, after some elementary calculations:

$$\begin{aligned} & [1 + kw_1(p)w_3(p) + kw_1(p)w_2(p)]x_{\text{output}}(p) \\ & = kw_1(p)w_2(p)x_{\text{input}}(p) \\ & - kw_1(p)w_2^2(p)w_{\text{convert}}(p)w_3(p)F(p) \\ & + w_2(p)F(p) + kw_1(p)w_2(p)w_3(p)F(p) \dots \end{aligned} \quad (14)$$

Assume that the transfer function of the stabilizing device has been selected in such a way that the structure obtained assures stability where there exists an unlimited increase in the gain,  $k$ . Dividing eqn (14) by  $kw_1(p)$ , and assuming that  $kw_1(p)$  tends to infinity, one obtains, after some simplifications:

$$\begin{aligned} & [w_3(p) + w_2(p)]x_{\text{output}}(p) = w_2(p)x_{\text{input}}(p) \\ & + [w_2(p)w_3(p) - w_2^2(p)w_{\text{convert}}(p)w_3(p)]F(p) \dots \end{aligned} \quad (15)$$

As is evident from (15), in order to eliminate the effect of disturbances, the transfer function of the converting device should be selected from the condition:

$$w_2(p)w_3(p) - w_2^2(p)w_{\text{convert}}(p)w_3(p) = 0 \quad (16)$$

or:

$$w_{\text{convert}}(p) = 1/w_2(p)$$

The realization of a device with a transfer function (16) may be attained by methods of constructing structures that are stable in the presence of an unlimited increase in the gain<sup>6</sup>, and this involves neither fundamental nor technical difficulties.

Generally speaking, the elimination of the influence of disturbances, in the given case, could be accomplished by the method described by the author<sup>7</sup>. Naturally it is expedient to make use of the indicated method if there are no additional disturbances at the system's input. If, at the system's input, there are disturbances in addition to the useful signal, then it is possible to show that the solution given here is more noise-proof. Let us convince ourselves of the accuracy of this affirmation.

It is assumed that, in place of the transfer function  $kw_1(p)$ , and in place of a stabilizing device with a transfer function  $w_3(p)$ , which provides for stability in the presence of an unlimited gain  $k$ , a system having the form shown in *Figure 4* is realized.

The introduction of an amplifier, with a high gain, which is encompassed by a stabilizing device with a transfer function  $w'_3(p)$ , depends on the necessity of providing stability to the entire system in the presence of unlimited increase in  $k$ .

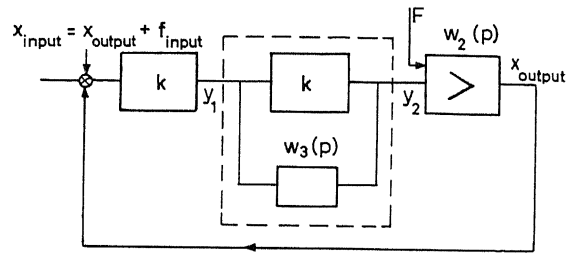


Figure 4

If  $w_2(p)$  has a power ' $p$ ' in the denominator that is greater than a 'fourth' one, then, as is shown<sup>6</sup>, it is possible to introduce several amplifiers with high gains and realize a structure that would admit an unlimited increase in  $k$  without disturbing the stability.

As is clear from Ref. 7 in this case, if  $x_{\text{input}}$  contains no disturbances, then an increase in the gain  $k$ , up to rather high values, eliminates the effect of the disturbances  $F$ .

Assume that the input signal contains a disturbance  $f_{\text{input}}$  to the extent that

$$x_{\text{input}} = x_{\text{input}u} + f_{\text{input}}$$

where  $x_{\text{input}u}$  is the disturbance-free input signal.

A system of equations for the circuit in *Figure 4* is drawn up, for the case under examination. At the same time, instead of the part of the circuit surrounded by a dotted line in *Figure 4*, assuming that here  $k$  is a sufficiently large number, one should straightway insert  $1/w'_3(p)$ .

$$Y_1(p) = k[x_{\text{input}u}(p) + f_{\text{input}}(p) - x_{\text{output}}(p)] \dots \quad (17)$$

$$\begin{aligned} Y_2(p) &= Y_1(p) \frac{1}{w'_3(p)} \\ &= \frac{k}{w'_3(p)} [x_{\text{input}u}(p) + f_{\text{input}}(p) - x_{\text{output}}(p)] \end{aligned} \quad (18)$$

$$x_{\text{output}}(p) = w_2(p) [Y_2(p) + F(p)] \quad (19)$$

\* Eqn (13):

$$Y(p) = \frac{kw_1(p)x_{\text{input}}(p) - kw_1(p)x_{\text{output}}(p) - kw_1(p)w_2(p)w_{\text{convert}}(p)w_3(p)F(p)}{1 + kw_1(p)w_3(p)} \quad (13)$$



Substituting the value  $Y_2(p)$  from (18) in eqn (19), one obtains:

$$x_{\text{output}}(p) = -\frac{k w_2(p) x_{\text{input}u}(p)}{w'_3(p)} + w_2(p) \frac{k f_{\text{input}}(p)}{w'_3(p)} + w_2(p) F(p) - \frac{k x_{\text{output}}(p)}{w'_3(p)} w_2(p)$$

or: 
$$\left[ 1 + \frac{k w_2(p)}{w'_3(p)} \right] x_{\text{output}}(p) = \frac{k w_2(p)}{w'_3(p)} [x_{\text{input}u}(p) + f_{\text{input}}(p)] + w_2(p) F(p)$$

When  $k$  tends to infinity, one obtains:

$$\frac{w_2(p)}{w'_3(p)} x_{\text{output}} = \frac{w_2(p)}{w'_3(p)} [x_{\text{input}u}(p) + f_{\text{input}}(p)]$$

or: 
$$x_{\text{output}}(p) = x_{\text{input}u}(p) + f_{\text{input}}(p) \quad (20)$$

Consequently, one obtains at the output a magnitude that is equal to the ideal input plus the disturbance.

Consider, at this point, the size of the magnitude at the output, in the presence of interference at the system's input, and with the elimination of the effect of disturbance  $F$  by the above-mentioned method.

Keeping in mind that at the input of the system in Figure 4, along with the useful signal, there is a disturbance input, one has the following system of equations (Figure 3)

Eqn (21) \*

or, considering (11), one has:

Eqn (22) †

from which:

Eqn (23) ‡

Substituting the value of  $y(p)$  from (23) in (9), and after some elementary calculations, one obtains:

Eqn (24) §

On fulfilling the condition  $w_2(p) = 1/w_3(p)$  and when  $k$  tends to infinity, one obtains:

$$x_{\text{output}}(p) = \frac{w_2(p)}{w_3(p) + w_2(p)} x_{\text{input}u}(p) + \frac{w_2(p)}{w_3(p) + w_2(p)} f_{\text{input}}(p) \quad (25)$$

By comparing the results expressed in eqn (20) and in eqn (25), one can draw the following conclusions. In the first case (eqn 20), the greater the gain, the closer the output magnitude to the sum of the ideal input plus the full disturbance at the input. In the second case (eqn 25), the picture is different. Depending on the properties of the useful signal and of the disturbance, especially for those cases where the frequency properties of the disturbance and the useful signal are different, the parameters  $w_3(p)$  may be selected in such a way as to reduce the interference, at the input, together with the useful signal, to a minimum.

#### A Change in the Plant's Parameters Taking Place as a Result of a Change in Operating Conditions or in Internal Factors

This case pertains to plants, in which, in the course of operation, the parameters of the plant itself may vary within a wide range. In such cases, the sensitivity factor, according to Bode<sup>8</sup>, represents an essential quality index of the entire system. For the plants being considered here, the sensitivity factor may be expressed in the following manner.

Assume that the plant's transfer function, as before, is designated by  $w_2(p)$ . The overall transfer function of the entire system in relation to changes in the plant being indicated by  $k(p)$ , the sensitivity, is expressed in the following way:

$$S_{w_2(p)}^k = \frac{\frac{dk(p)}{k(p)}}{\frac{dw_2(p)}{w_2(p)}} = \frac{dk(p)}{dw_2(p)} \cdot \frac{w_2(p)}{k(p)} \quad (26)$$

In the general case, the smaller the magnitude of  $S_{w_2(p)}^k$ , the less sensitive are the dynamic properties of the system, in its entirety, to changes in the plant's properties. For this case, the system is considered ideal or self-adjusting, if the magnitude  $S_{w_2(p)}^k$  does not depend on the characteristics of  $w_2(p)$  or  $S_{w_2(p)}^k$  tends to 0.

The following proof is given. Structures that are stable in face of an unlimited increase of gain, in which stability is

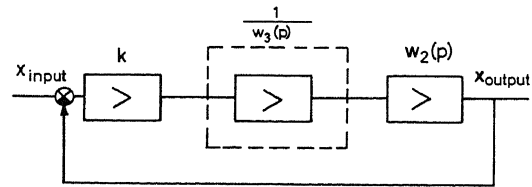


Figure 5

\* Eqn (21):  $Y(p) = k w_1(p) \{x_{\text{input}u}(p) + f_{\text{input}}(p) - x_{\text{output}}(p) - w_3(p) [y(p) + (x_{\text{output}}(p) - x'_{\text{output}}(p)) w_{\text{convert}}(p)]\}$  (21)

† Eqn (22):  $Y(p) = k w_1(p) [x_{\text{input}u}(p) - f_{\text{input}}(p) - x_{\text{output}}(p) - w_3(p) y(p) - w_3(p) w_{\text{convert}}(p) w_2(p) F(p)]$  (22)

‡ Eqn (23):  $Y(p) = \frac{k w_1(p) [x_{\text{input}u}(p) + f_{\text{input}}(p) - x_{\text{output}}(p) - k w_1(p) w_2(p) w_3(p) w_{\text{convert}}(p) F]}{1 + k w_1(p) w_3(p)}$  (23)

§ Eqn (24):

$$x_{\text{output}}(p) = \frac{k w_1(p) w_2(p) x_{\text{input}u}(p) + k w_1(p) + f_{\text{input}}(p) - k w_1(p) w_2^2(p) w_{\text{convert}}(p) w_3(p) F(p) + k w_1(p) w_2(p) w_3(p) F(p)}{1 + k w_1(p) w_3(p) + k w_1(p) w_2(p)} \quad (24)$$

achieved by the introduction of ideal derivatives and whose degree of ideality is determined by the magnitude of the gain<sup>7</sup>, belong to the category of self-adjusting systems in the sense indicated above. There is no question about that. In *Figure 5* one observes the structure of a control system of the type under consideration. The transfer function of the closed loop will be written in the following form:

$$k(p) = \frac{\frac{k}{w_3(p)} w(p)}{1 + \frac{k}{w_3(p)} w_2(p)} \quad (27)$$

Now the expression for the sensitivity is found. In conformity with (26):

$$S_{w_2(p)}^k = \frac{\frac{k}{w_3(p)} \left[ 1 + \frac{k}{w_3(p)} w_2(p) \right] - \left( \frac{k}{w_3(p)} \right)^2 w_2(p)}{\left[ 1 + \frac{k}{w_3(p)} w_2(p) \right]^2} \cdot \frac{w_2(p) \left( 1 + \frac{k w_2(p)}{w_3(p)} \right)}{\frac{k}{w_3(p)} w_2(p)}$$

or, after simplification:

$$S_{w_2(p)}^k = \frac{1}{1 + \frac{k}{w_3(p)} w_2(p)} \quad (28)$$

When  $k$  tends to infinity,  $S_{w_2(p)}^k$  tends to 0. In other words, in the sense indicated above, one obtains an ideal system.

Now consider the expression for sensitivity, if the structure belongs to the category of those that are stable in the face of an unlimited increase of gain, and where stability is achieved by the introduction of passive stabilizing devices.

As an example, one should consider the simplest type of such a system whose structure is shown in *Figure 6* (excluding the dotted line part).

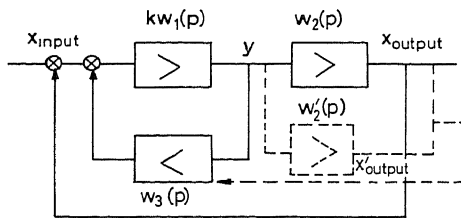


Figure 6

The transfer function of the closed loop in *Figure 6* is written as follows:

$$k(p) = \frac{k w_1(p) w_2(p)}{1 + k w_1(p) w_3(p) + k w_1(p) w_2(p)} \quad (29)$$

The sensitivity, according to  $w_2(p)$ , is written:

$$\boxed{\text{Eqn (30)}}^*$$

or, after simplification:

$$\boxed{\text{Eqn (30a)}}^\dagger$$

When  $k$  tends to infinity, one has:

$$\lim_{k \rightarrow \infty} S_{w_2(p)}^k = \frac{w_3(p)}{w_2(p) + w_3(p)} \quad (31)$$

Consequently, in the given case, even with sufficiently high gains, a change in the parameters or characteristics of the plant exerts an influence on the dynamic properties of the system.

Consider some methods for improving the system's structure, with the object of reducing to the minimum the effect of the variation in the plant's characteristics on the system's dynamic properties, and in this manner, make the system self-adjusting in the above-determined sense.

When external disturbances, which did not seem capable of measurement, act on the plant, it is expedient in this case, too, to introduce a plant model into the system, in order to obtain a self-adjusting system. A structural block diagram for the case under consideration is shown in *Figure 6* (including the dotted line part).

Keeping in mind the designations set forth in *Figure 6*, one writes:

$$\boxed{\text{Eqn (32)}}^\ddagger$$

Here,  $x'_{\text{output}}(p)$  is the representation of the output of the plant's model and  $x_{\text{output}}$  is the representation of the plant's output.

It is assumed that the model's characteristics remain invariable. Under these conditions, the difference  $x'_{\text{output}}(p) - x_{\text{output}}(p)$  is equivalent to the disturbance which depends on the change in the plant's characteristics. Consequently

$$x_{\text{output}}(p) - x'_{\text{output}}(p) = CF(p) \quad (33)$$

$C$  is a constant coefficient.

In this manner,

$$x_{\text{output}}(p) = w_2(p) y(p) = w_2'(p) y(p) + C w_2'(p) F(p) \quad (34)$$

\* *Eqn (30):*

$$S_{w_2(p)}^k = \frac{k w_1(p) [k w_1(p) w_3(p) + k w_1(p) w_2(p) + 1] - k w_1(p) k w_1(p) w_2(p)}{[1 + k w_1(p) w_3(p) + k w_1(p) w_2(p)]} \quad (30)$$

† *Eqn (30a):*

$$S_{w_2(p)}^k = \frac{k w_1(p) \cdot k w_1(p) w_3(p) + k w_1(p)}{k w_1(p) [1 + k w_1(p) w_3(p) + k w_1(p) w_2(p)]} \frac{k w_1(p) w_3(p)}{1 + k w_1(p) w_3(p) + k w_1(p) w_2(p)} \quad (30a)$$

‡ *Eqn (32):*

$$Y(p) = k w_1(p) [x_{\text{input}}(p) - x_{\text{output}} - w_3(p) y - w_3(p) (x_{\text{output}}(p) - x'_{\text{output}}(p))] \quad (32)$$

Substituting in eqn (32), instead of  $x'_{\text{output}}(p) - x_{\text{output}}(p)$ , the difference value from (33), one obtains:

$$\boxed{\text{Eqn (35)}}^*$$

From the above, the expression for  $Y(p)$  is found:

$$Y(p) = \frac{k w_1(p) x_{\text{input}}(p) - k w_1(p) x_{\text{output}}(p) - k w_1(p) w_3(p) CF(p)}{1 + k w_1(p) w_3(p)} \quad (36)$$

Substituting the value for  $y(p)$  from (36) in eqn (34), one obtains:

$$\boxed{\text{Eqn (37)}}^\dagger$$

or, determining  $x_{\text{output}}(p)$  from (37), one obtains:

$$x_{\text{output}}(p) = \frac{k w_1(p) w'_2(p) x_{\text{input}} + w'_2(p) CF(p)}{1 + k w_1(p) w'_2(p) + k w_1(p) w_3(p)} \quad (38)$$

When  $k$  tends to infinity:

$$x_{\text{output}}(p) = \frac{w_1(p) w'_2(p) x_{\text{input}}(p)}{w_1(p) w_3(p) + w_1(p) w'_2(p)} = \frac{w'_2(p) x_{\text{input}}(p)}{w_3(p) + w'_2(p)} \quad (39)$$

From (39) it is evident that the output magnitude does not depend on the change in parameters of the controlling device. Under the conditions where  $w'(p)$  corresponds to the optimum operating circumstances, from the point of view of some quality criterion, the process in the system will be maintained automatically at these working conditions, independently of the plant's characteristic changes.

Thus, as a result of considering the three most interesting

cases involving changes in the characteristics of the controlled plants—changes due to the effect of external disturbances, which could be measured, those due to external disturbances that do not appear to be capable of measurement, and those which result from plant characteristic changes in the course of operation that are independent of external disturbances—methods are suggested for designing structures which would provide for the independence of the plant's selected operating conditions from possible external and internal effects on it, and, consequently, the structures obtained prove to be self-adjusting system structures.

## References

- 1 FELDBAUM, A. A. On the use of computers in automatic systems. *Automat. Telemekh., Moscow* No. 11 (1956)
- 2 FELDBAUM, A. A. Automatic optimizer. *Automat. Telemekh., Moscow* No. 8 (1958)
- 3 DOGANOVSKI, S. A., and FELDBAUM, A. A. Study of compensation for carrier thickness oscillations with the aid of an electron model. *Automat. Telemekh., Moscow* No. 2 (1959)
- 4 FELDBAUM, A. A. *Computers in Automatic Systems*. 1959. Moscow; Fizmatgiz
- 5 LERNER, A. YA. Optimum high-speed control system by means of an automatic potentiometer. *Automat. Telemekh., Moscow* No. 2 (1952)
- 6 MEEROV, M. V. *Structure Synthesis of Highly-accurate Automatic Control Systems*. 1959. Moscow; Fizmatgiz
- 7 MEEROV, M. V. On the structural noiseproof feature of one category of dynamic systems. *DAN, Proc. Acad. Sci.* No. 4 (1961)
- 8 *Chain Theory and Construction of Return-communication Feedback Amplifiers*. 1948. Moscow; Fizmatgiz

$$* \text{ Eqn (35):} \quad Y(p) = k w_1(p) [x_{\text{input}}(p) - x_{\text{output}}(p) - w_3(p) y(p) - w_3(p) CF(p)] \quad (35)$$

$$\dagger \text{ Eqn (37):} \quad x_{\text{output}} = \frac{w'_2(p) [k w_1(p) x_{\text{input}}(p) - k w_1(p) x_{\text{output}}(p) - w_3(p) k w_1(p) CF(p)]}{1 + k w_1(p) w_3(p)} + w'_2(p) CF(p) \quad (37)$$

## DISCUSSION

D. A. BELL, *A.M.F. British Research Laboratory, Reading, Berks., England*

While I agree with most of Professor Meerov's conclusions, which have something in common with the work of Horowitz<sup>1</sup>, I have doubts about the effects of gross non-linearities. A moderate degree of non-linearity can be handled by the method of describing functions, but if there is complete saturation, I believe this invalidates the whole concept of (linear differential equation) transfer functions in terms of which the whole of the author's work is presented. Is this point discussed in Reference 6, and has that reference been translated into languages other than Russian?

I do not agree that an ideal servo system is *ipso facto* a 'self-adjusting' servo. I suggest first that a servo system is a closed-loop system, and that 'closed loop' implies normally the existence of unidirectional paths and a finite time of propagation of signals round the loop. (This distinguishes a closed-loop system from an action-reaction situation.) Such a closed-loop system has a transfer function, and I would define an adaptive servo, or one of which the transfer function is automatically modified, as a result of observation of its performance.

The author refers in connection with eqn (25) to the noise-filtering properties of a servo, and I believe the major use for adaptive servo

systems is where the statistical characteristics of either signal or noise are liable to change.

The other limitation of the author's ideal systems is that they require infinite gain-bandwidth in part, at least, of the loop, bearing in mind the restrictions imposed by signal level and linearity requirements. I doubt whether virtually infinite gain-bandwidth is always attainable. An adaptive system may then be needed to shift the available gain-bandwidth of the control amplifier into the band in which it is currently required by a varying controlled system.

## Reference

- 1 HOROWITZ, J. M. Plant adaptive systems versus ordinary feedback systems. *Trans. Inst. Radio. Engrs.* AC-7, No. 1. (1962) 48

J. B. CRUZ, *University of Illinois, Champaign, Illinois, U.S.A.*

The structures proposed by the author are quite interesting. Although the derivations of the expressions for the output quantities for the structures proposed by the author are correct, all the results could be written down in one step if Mason's<sup>4</sup> rule for signal flow graphs is used. In particular, one could avoid the intermediate eqns (3), (4) and (5),

to obtain eqn (6); eqns (8), (9), (10), (11), (12) and (13) leading to eqn (14) as well as eqns (17), (18), (19) and other (unnumbered) equations after (19) could be deleted. Likewise, eqns (21), (22), (23) and (24) which are two columns wide, could be avoided. It appears that using Mason's rule, the length of the paper could be reduced by a factor of 2.

The second case considered by the author, where the disturbance input to the plant cannot be measured directly, and the third case, where there are parameter changes in the plant, are very directly related. The structure of Figure 3 which is the one proposed for case 2 as well as case 3 (except that  $w_{\text{convert}}(p)$  is absent) is reducible to a typical two-degree-of-freedom structure considered by Horowitz<sup>2</sup>. It is well known that for two-degree-of-freedom systems such as those used by Professor Meerov in cases 2 and 3, the problem of minimizing the effects of disturbance inputs is not independent of the problem of minimizing the effects of plant parameter variations. As a matter of fact, if the loop gain is made to approach infinity, the sensitivity to parameter variations goes to zero and the effect of plant disturbance inputs also goes to zero<sup>3</sup>. However, it is tacitly assumed that the system remains stable. If the equivalent canonic two-degree-of-freedom structure becomes unstable as the loop gain is increased, then the structure proposed by the author could be unstable since the transfer function is not changed. Making the loop gain as high as stability will allow is, of course, well known in feedback system design, if low sensitivity is a major objective.

The third comment I make is with respect to the assumption that all controllers are realized exactly and amplifier gain can be made infinite. Since it is not obvious whether deviations from these assumptions correspond to second-order effects only, I ask Professor Meerov if he has pursued these matters further.

#### References

- MASON, S. J. Signal flow graphs — further properties. *Proc. Inst. Radio Engrs.* (July 1956)
- HOROWITZ, I. M. *Synthesis of Feedback Systems*. 1963. New York; Academic Press

G. SCHMIDT, *Institut für Regelungstechnik, Schloßgraben 1, Technische Hochschule Darmstadt, W. Germany*

My remark concerns the last part of the paper. All the control systems considered by the author may be reduced to the single loop two-degree-of-freedom structures investigated by Horowitz<sup>1</sup> and others. There exists a very simple connection between the sensitivity function and the control elements.

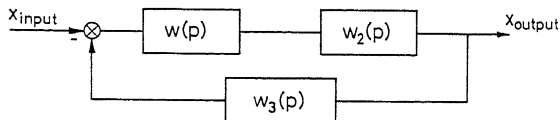


Figure A

The transfer function of the system shown in Figure A is

$$k(p) = \frac{x_{\text{output}}(p)}{x_{\text{input}}(p)} = \frac{w_1(p) w_2(p)}{1 + w_1(p) w_2(p) w_3(p)}$$

From this expression one can find by differentiation, and after some simplification, the sensitivity function corresponding to  $w_2(p)$ :

$$S_{w_2}^k(p) = \frac{\frac{dk(p)}{dw_2(p)}}{k(p)} = \frac{1}{1 + w_1(p) w_2(p) w_3(p)}$$

It is evident that the denominator of this sensitivity function contains

only the open-loop transfer function of the control system. Consequently eqns (28) and (31) of the paper may be written in a much simpler way than was done by the author.

#### Reference

- HOROWITZ, J. M. *Synthesis of Feedback Systems*. 1963. New York; Academic Press

J. WILLIAMS, *Engineering Laboratory, Oxford University, England*

Feedback systems are, by their very nature, 'self-adaptive' and the fundamental reason for using feedback in the first instance was that the output of such a system would be relatively insensitive to parameter changes. To illustrate that one can obtain an invariant system with feedback: consider Figure A.

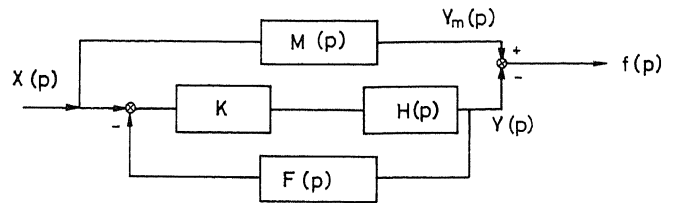


Figure A

Let  $M(p)$  represent the transfer function of a model which has been designed to give a desired performance under all conditions of operation.  $H(p)$  is the transfer function of the plant and may include stabilizing networks so that large values of the amplification factor  $K$  may be used.  $F(p)$  is the transfer function of the feedback network.

We require:  $Z(p) = Y_m(p) - Y(p) \equiv 0$

Substituting for  $Y_m(p)$  and  $Y(p)$

$$Z(p) = X(p) \left[ M(p) - \frac{K \cdot H(p)}{1 + KH(p)F(p)} \right]$$

For large  $K$ :

$$Z(p) = X(p) \left[ M(p) - \frac{1}{F(p)} \right]$$

For  $Z(p) = 0$ :

$$M(p)F(p) = 1$$

Therefore if the feedback network is made inverse of the model, and provided  $K$  is sufficiently large, the output of the system will be equal to the model (desired) output.

M. V. MEEROV, *in reply*

To begin with, I single out the remarks of Mr. Williams who expressed the view that, generally, a system with feedback is, by its nature, adaptive. This is true, but in modern automatic control theory and practice the concept of self-adjustment has a somewhat different sense.

Although, as will be seen below (and it seems to me that this follows from my paper and that of Mr. Kochenburger), for a system to be adaptive it is not necessary that it contains search elements, it must nevertheless possess something more than an ordinary system with feedback.

The other three comments—by J. B. Cruz, D. A. Bell and G. Schmidt—concern the following points:

(1) It is assumed in my paper that the system remains stable with an increase in the gain. The point raised is how this is and what has been published on the theory of construction of systems which are stable with unlimited increase in gain.

(2) The question of non-linearity.

(3) The connection between the results in my paper and those of J. M. Horowitz.

(4) Are the systems described in my paper adaptive?

In 1947<sup>1</sup> I published a paper proving the existence of a class of structures for which there was no contradiction between stability and precision. In systems with such structures, it is possible to increase the gain without constraint, without impairing stability. Methods were derived for building such structures and their principal dynamic properties ascertained.

A number of other studies in this direction were subsequently published. My book<sup>2</sup> summed up all the previous studies. The ideas in the paper are based mainly on the results of this book, which also examines questions connected with the evaluation of the influence of certain non-linearities. Owing to lack of space I was, of course, unable in the paper to make more than brief references to the material in the book. However, this book is now being translated into English, and will be published shortly. I would add that some of the relevant material may also be read elsewhere<sup>3</sup>. These references should answer the majority of these points.

On the third question the answer is straightforward. It is shown in the paper that among the structures which are stable, however great gain, there is a subclass for which  $-\int_{\omega}^K$  (sensitivity according to Baudet) automatically (structurally) tends to zero.

Horowitz has shown that to make  $\int_{\omega}^K$  small, the structure must be two-dimensional. The difference between my work and that of Horowitz is therefore obvious.

In reply to the final question, an illustration shows the practical significance of the structures obtained.

No one will doubt that the system proposed by Kochenburger is adaptive, so I shall give the solution of Kochenburger's problem by my method.

There is a plant with transfer function  $\frac{K}{(1+0.2p)^2(1+0.005p)}$  where, in the process of operation,  $K$  can change 100 times, and, moreover, quite rapidly.

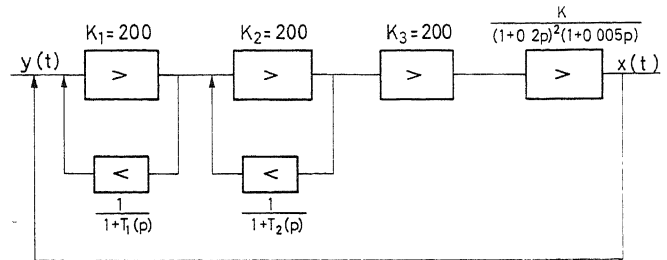


Figure A

Kochenburger's solution is well known, being contained in his paper read at this I.F.A.C. Congress. Figure A shows the layout and data of the individual elements of the circuit which solves the same problem by our method,  $T_1 = T_2 = 0.01$  sec.

Figure B shows the oscillogram for  $K = 10$ , and Figure C the oscillogram for  $K = 0.1$ , i.e. 100 times smaller. Figure D gives the oscillogram for  $K = 5 - 4 \sin 6.28 t$ .

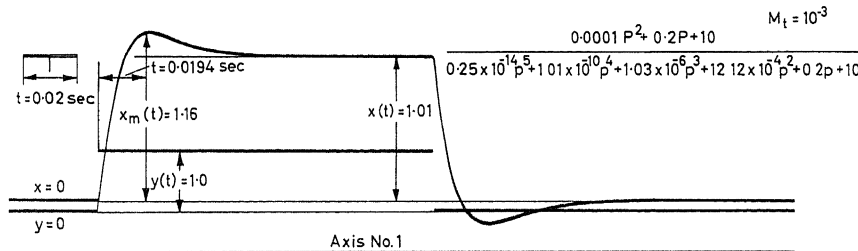


Figure B

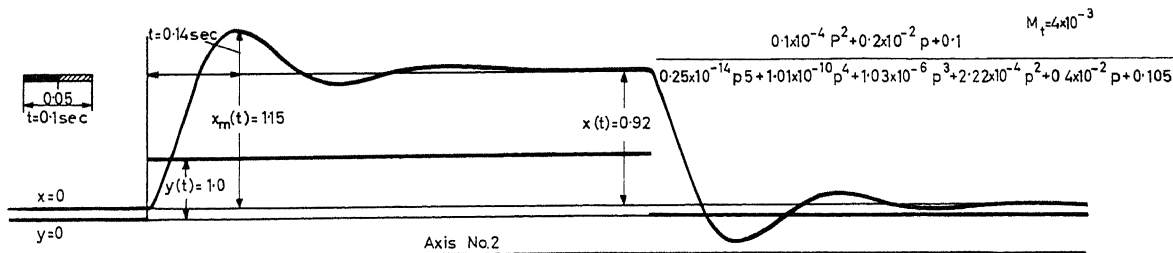


Figure C

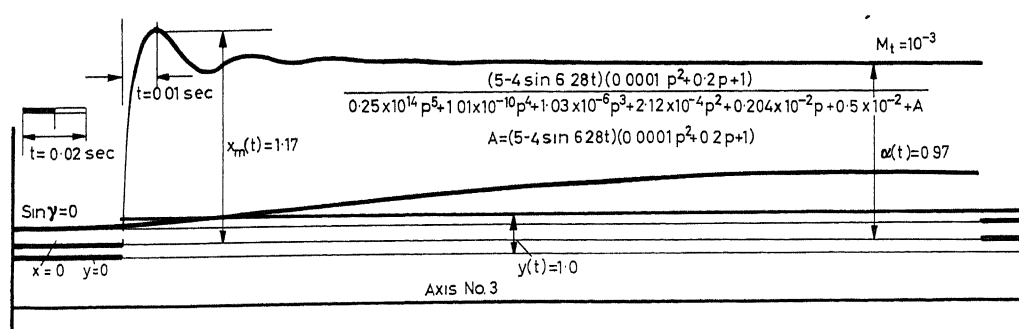


Figure D

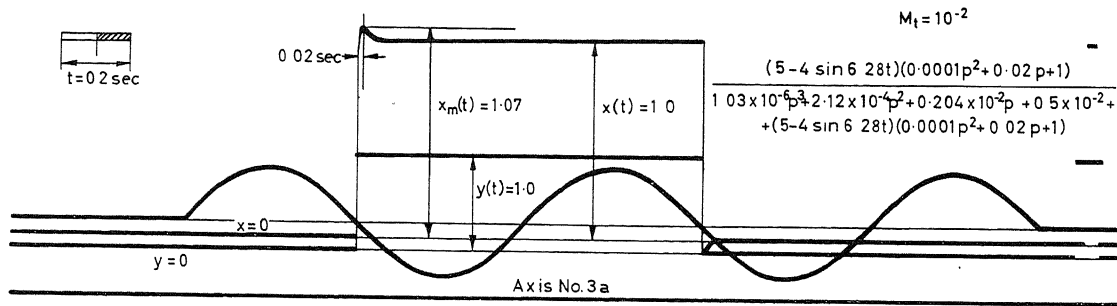


Figure E

Figure E shows an oscillogram for the same data, but with the gains of the first two amplifiers doubled, i.e.  $K_1 = 400$ ,  $K_2 = 400$ . As can be seen from these oscillograms, the output quantity  $X(t)$  for  $Y(t) = \text{const.}$  remains constant irrespective of the variations in the plant gain.

I doubt that anyone could consider that this system is not equivalent to an adaptive system, merely because it is simple and does not contain complex search elements or any computing devices.

#### References

- <sup>1</sup> MEEROV, M. V. *Avtomat. Telemekh.* No. 4 (1947)
- <sup>2</sup> MEEROV, M. V. *Synthesis of Structures of High-Precision Automatic Control Systems.* 1959. Moscow; Fizmatgiz
- <sup>3</sup> MEEROV, M. V. *Introduction to the Dynamics of Automatic Control of Electrical Machines.* Chaps. 9 and 11. 1956. Moscow; Academy of Sciences; 1960. London; Butterworths

# A Comparison of the Measuring Time in Self-adjusting Control Systems

F. MESCH

## Summary

It seems impossible to select the 'best' self-adjusting system out of the many proposals, but it is possible to compare them from a certain point of view. In this paper the point of view is the time required in measuring the parameters to be adjusted, considering random input signals. Three devices are compared which adjust one parameter of a control system in such a way that the R. M. S. error becomes a minimum.

Instead of general derivations the comparison is demonstrated, using a very simple control system as an example. In each case the variance of the measuring unit output is estimated, this varies statistically because of the finite averaging time. From the variance the necessary time constants of the smoothing filter are determined.

**Results:** From the three devices the 'stepwise' searching method requires long measuring intervals. The method with a 'periodic searching signal' is better because the search by trial and error is replaced by an adjustment directed uniquely to the minimum, but the measuring interval is in the same order of magnitude. The method with a 'reference model', however, has also the advantage of a shorter measuring time.

## Sommaire

S'il est impossible de sélectionner le meilleur des systèmes auto-régulateurs proposés, on peut au moins les comparer d'après un certain critère. Dans ce rapport, le critère choisi est le temps de mesure nécessaire des paramètres réglés, en supposant que les signaux d'entrée soient aléatoires. Avec ce critère, on compare trois éléments qui ont tous pour but de régler un paramètre dans un système de commande en minimisant l'erreur moyenne quadratique. Pour effectuer la comparaison, on utilise un système de commande très simple. Dans chaque cas, on estime la variance de la sortie de l'élément de mesure; cette variance varie statistiquement en raison du temps fini d'établissement de la moyenne. A partir de la variance, on détermine les constantes de temps nécessaires du filtre de lissage.

**Résultats de comparaison:** La méthode de recherche «par paliers» demande de longues durées de mesure. La méthode avec un «signal de recherche périodique» est meilleure, parce que le tâtonnement est remplacé par un réglage uniquement orienté vers le minimum, mais la durée de la mesure est du même ordre de grandeur. La méthode avec un «modèle de référence» a un temps de mesure plus court.

## Zusammenfassung

Es ist kaum möglich, aus den vielen vorgeschlagenen selbststellenden Systemen das „beste“ auszuwählen. Möglich ist der Vergleich nach bestimmten Gesichtspunkten. In der vorliegenden Arbeit ist dieser Gesichtspunkt die Meßzeit, die nötig ist, um bei regellosen Eingangssignalen die einzustellenden Parameter zu messen. Verglichen werden drei Anordnungen, die in einem Regelkreis einen Parameter selbstständig so einstellen, daß die mittlere quadratische Regelabweichung zu einem Minimum wird.

Statt allgemeiner Ableitungen wird der Vergleich an einem sehr einfachen Regelkreis als Beispiel demonstriert. Dazu wird jedesmal die Streuung der Ausgangsgröße der Meßeinrichtung abgeschätzt, die wegen der endlichen Meßzeit regellos um den gesuchten Wert schwankt. Hieraus ergeben sich die notwendigen Zeitkonstanten des Glättungs-filters.

**Ergebnisse:** Von den drei Anordnungen erfordert das „schrittweise Suchverfahren“ lange Meßzeiten. Das Verfahren mit „periodischem Suchsignal“ ist insofern besser, als das probierende Suchen durch eine zielrichtige Verstellung des Parameters ersetzt wird; die Meßzeit hat aber die gleiche Größenordnung. Das Verfahren mit einem „Bezugsmodell“ dagegen hat zielrichtige Verstellung und kürzere Meßzeiten.

## Introduction

During recent years many different systems have been proposed which automatically adjust one or more of their dynamic parameters. Since these are always non-linear systems it seems impossible to compare the proposals in a general manner and to select the 'best' of them, because no accepted quality measure exists.

In the presence of random input signals which normally occur in control systems, the time needed for the measurement of the parameters to be adjusted may, as will be shown, gain considerable importance. Some authors did not pay enough attention to this point in making their proposals. The measuring time will therefore be the point of view from which some typical systems are compared.

## A Control System as an Example

The comparison is demonstrated using the example of a very simple control system, as shown in the upper portion of Figure 1. The only free element in it is the gain  $K_R$  of the controller.

The input signal  $w(t)$  and the disturbance  $z(t)$  are assumed to have the power spectral densities  $S_{ww} = k^2/\omega^2$  and  $S_{zz} = \kappa^2 k^2$ , respectively. The parameters  $k$ ,  $\kappa$  and the gain  $K_S$  of the process may change in an unpredictable manner. The optimal adjustment of  $K_R$  is to be found, using the mean square of the error

$$\overline{x_w^2(t)} = \lim_{T \rightarrow \infty} \frac{1}{2T} \int_{-T}^T x_w^2(t) dt \quad (1)$$

as an optimization criterion or performance index.

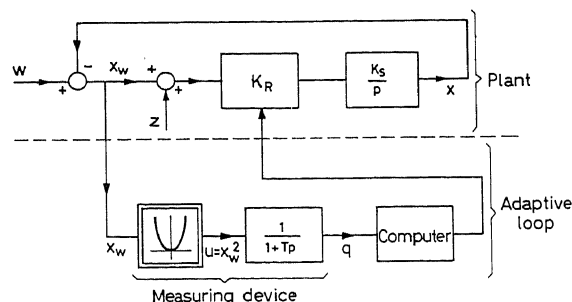


Figure 1. Stepwise searching method

This mean square value may also be calculated in a common way from the power spectral density of the error  $x_w$  as a mean power by

$$\overline{x_w^2(t)} = \int_0^\infty S_{x_w x_w}(\omega) d\omega \quad (2)$$

providing stationary input signals and constant parameters. Since the parameters, however, are subject to changes, the assumption is made for the following that they are changing slowly compared with the relevant time constants of the system; so it is still possible to consider the processes as stationary and to use eqn (2).

Applying this to the control system of *Figure 1* and introducing  $K = K_R \cdot K_S$ ,

$$S_{x_w x_w}(\omega) = \frac{\omega^2}{K^2 + \omega^2} \cdot S_{w w}(\omega) + \frac{K^2}{K^2 + \omega^2} \cdot S_{z z}(\omega) \quad (3)$$

With the assumed spectra  $S_{w w}$  and  $S_{z z}$  one obtains

$$S_{x_w x_w}(\omega) = \frac{k^2(1 + K^2 \kappa^2)}{K^2 + \omega^2}$$

and

$$\overline{x_w^2(t)} = \frac{\pi}{2} k^2 \left( \frac{1}{K} + \kappa^2 K \right) \quad (4)$$

where a minimum is found at

$$K_{\text{opt}} = \frac{1}{\kappa} \quad (5)$$

### Stepwise Searching Method

The most obvious idea for the automatic search of this minimum is to make a computer measure the criterion for various values of  $K$  successively and then select the best one. Therefore, the quantity  $Q$  to be measured by the adaptive loop is the mean square error itself, giving  $Q = \overline{x_w^2(t)}$ , and the corresponding measuring device is shown in the lower portion of *Figure 1*. The integration over an infinite time interval according to eqn (1) is impossible, of course; a finite time integration would have to be repeated permanently for the self-adjustment. An easier realizable averaging device than an ideal integrator is a continuously acting low-pass filter, consisting of a simple lag with time constant  $T$ . Because of the finite averaging time, there does not occur the desired value  $Q$  at the output of the low-pass filter but a random variable  $q(t)$  fluctuating about the true value  $Q$ :

$$q(t) = Q + q_1(t)$$

where  $q_1 = q - Q = q - \bar{q}$  denotes the fluctuating (a.c.) component. The magnitude of the fluctuation obviously depends on the filter time constant  $T$ . To compare  $T$  of different systems, this dependence requires a more detailed analysis.

The magnitude of the fluctuation can be expressed as usual by the standard deviation  $\sigma_q$  of  $q$ . One obtains  $\sigma_q^2$  either from the autocorrelation function  $\phi_{qq}(\tau)$  for  $\tau = 0$  (see for example references 1 and 2) or from the spectrum  $S_{qq}(\omega)$  by integration<sup>3</sup> over  $\omega$ .

Here the latter way is chosen. At first one calculates the spectrum  $S_{uu}(\omega)$  behind the squarer<sup>4, 5</sup>:

$$S_{uu}(\omega) = 2 \phi_{x_w x_w}^2(0) \cdot \delta(\omega) + \int_{-\infty}^{\infty} S_{x_w x_w}(\omega_1) \cdot S_{x_w x_w}(\omega - \omega_1) d\omega_1 \quad (6a)$$

where  $\phi_{x_w x_w}^2(0) = (\overline{x_w^2(t)})^2 = Q^2$  is the mean value of  $u(t)$  and the integral on the right hand represents the spectrum  $S_{1_{uu}}(\omega)$  of the fluctuating component.

For the example this becomes by eqn (3)

$$S_{1_{uu}}(\omega) = \frac{2 \pi k^4 (1 + K^2 \kappa^2)^2}{K} \cdot \frac{1}{4 K^2 + \omega^2} \quad (6b)$$

and from the spectral density of the fluctuating component at the low-pass filter output

$$S_{1_{qq}}(\omega) = \frac{1}{1 + \omega^2 T^2} \cdot S_{1_{uu}}(\omega)$$

one obtains after integration over  $\omega$

$$\sigma_q^2 = \int_0^\infty S_{1_{qq}}(\omega) d\omega = \frac{\pi^2 k^4 (1 + K^2 \kappa^2)^2}{2 K^2 (1 + 2 K T)}$$

As shown below, the smoothing time constant  $T$  is much larger than the time constant  $1/K$  of the closed loop, thus  $KT \gg 1$ ; relating the variance  $\sigma_q$  to the exact value  $Q$ , one obtains

$$\frac{\sigma_q}{Q} \approx \frac{1}{(KT)^{\frac{1}{2}}} \quad (7)$$

Because of the non-linear operation of 'squaring', the amplitude distribution of  $q$  is not normal. However, if  $T$  becomes sufficiently large and thus  $\sigma_q \ll Q$ , the distribution, due to the central limit theorem of probability, approaches a Gaussian one with 95 per cent of all values lying within  $\pm 2\sigma$ . Setting the desired accuracy

$$\frac{P\%}{100} = \frac{2\sigma_q}{Q}$$

the actual value will be within the error limit  $\pm P \cdot Q/100$  with a 95 per cent probability. This results in the smoothing time constant

$$T = \frac{4 \cdot 10^4}{(P\%)^2 \cdot K} \quad (8a)$$

and for the desired accuracy of 5 per cent

$$T = 1,600/K \quad (8b)$$

that is, 1,600 times as large as the time constant of the control loop. Since several trial steps are usually necessary to reach the optimal adjustment, the result is a very large, often intolerable, amount of time.

### Sinusoidal Searching Signal

In optimizing controls there is a well-known technique of measuring the slope of the curve, the extremum of which is searched, by using a small periodic perturbation signal and a subsequent phase-sensitive detection. Instead of a physical variable, it is also possible of course to perturb a dynamic parameter periodically, and from *Figure 1* follows the system of *Figure 2*<sup>6, 7</sup>. In the example  $K$  is changed sinusoidally:

$$K' = K(1 + m_1 \cdot \sin \omega_s t) \quad (9)$$

The mean square error, obtained as in *Figure 1*, is being multiplied by the same searching signal (phase-sensitive detection)



and smoothed in low-pass filter 2. The low-pass filter 1 (dotted lines) can then be saved, and the following considers only the one filter 2.

As already seen, the measurement of a parameter requires time. The periodical perturbation of such a parameter must therefore be very slow, which means that  $\omega_s$  must be small compared to the reciprocal of the time constant of the control loop. This assumption permits one to calculate the mean value  $\bar{u} = \bar{x_w^2}$  of the squaring device output  $u(t)$ , as on the first page of this paper, and to subsequently introduce the sinusoidal parameter perturbation ( $\bar{u}$  denotes the averaging over a time interval that is small in comparison to the period of the searching signal). Thus eqns (4) and (9) give for this example

$$\bar{x_w^2} = \frac{\pi}{2} k \left[ \frac{1}{K(1+m_1 \sin \omega_s t)} + \kappa^2 K(1+m_1 \sin \omega_s t) \right]$$

After multiplication by  $s = m_2 \sin \omega_s t$  and final smoothing, the quantity  $Q$  measured by the adaptive loop now becomes, with  $m_1 \ll 1$ ,

$$Q = \bar{x_w^2(t)} \cdot m_2 \sin \omega_s t \approx \frac{\pi}{4} k^2 m_1 m_2 \left( \kappa^2 K - \frac{1}{K} \right) \quad (10)$$

the self-adjusting loop being supposed opened behind the measuring unit. As desired, the quantity  $Q$  is proportional to the differential quotient of the criterion  $x_w^2 = f(K)$  and zero at the minimum  $K_{opt} = 1/\kappa$ . At both sides of the extremum,  $Q$  has different signs and thus behaves like the error of a conventional control system. This enables the trial-and-error method to be replaced by a methodical adjustment of the parameter, and a simple on-off controller is sufficient, as indicated in Figure 2.

This is an important advantage of the method. However, regarding the measuring time it offers no progress. For proving, the variance  $\sigma_q$  of  $q$  related to  $Q$  is estimated in Appendix I to be

$$\frac{\sigma_q}{Q} \approx \frac{\sqrt{2}(K^2 \kappa^2 + 1)}{(K^2 \kappa^2 - 1) \cdot m_1} \cdot \frac{1}{(KT)^{\frac{1}{2}}} \quad (11a)$$

Setting, for instance,  $K = 2/\kappa$ , that is twice as large as optimal, and  $m_1 = 0.1$  corresponding to a perturbation of  $K$  by 10 per cent,

$$\frac{\sigma_q}{Q} \approx \frac{24}{(KT)^{\frac{1}{2}}} \quad (11b)$$

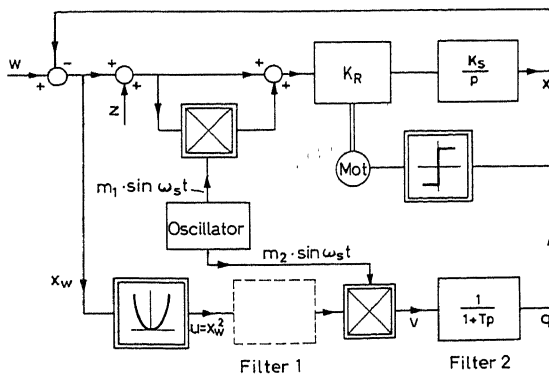


Figure 2. Sinusoidal searching signal

This seems at first to be still worse than eqn (7) for the stepwise searching method; the reason is that the quantity  $Q$ , with approximately the same  $\sigma_q$ , is smaller by the modulation index  $m_1$ . But it must be taken into account that here the accuracy of the measurement is not as far as critical; only the sign of  $q$  must be correct for the parameter adjustment running towards the optimum. Hence  $2\sigma_q = Q$  is sufficient (see Figure 3), and with the same numerical values as in eqn (11b) one obtains

$$T = \frac{8(K^2 \kappa^2 + 1)^2}{m_1^2 (K^2 \kappa^2 - 1)^2 \cdot K} = \frac{2,200}{K} \quad (12)$$

This is the same order of magnitude as with the stepwise searching method.

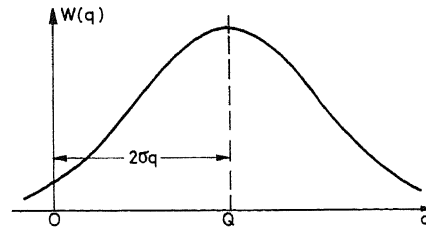


Figure 3. Amplitude distribution density of the measured variable  $q$

The modulation index  $m_1$  has a quadratic influence on  $T$ , and increasing  $m_1$  could save measuring time. The selected value  $m_1 = 0.1$ , however, might be adequate for practical applications; otherwise the desired approach to the optimum would be frustrated by excessive searching movements.

At the optimum itself, where  $K \cdot \kappa = 1$  and  $Q = 0$ ,  $\sigma_q/Q$  and  $T$  become infinite. One will make a compromise here, too, between the measuring time and the accuracy of optimization.

### Model with Parameter-perturbation

The stepwise searching method (shown previously), measures the performance criterion in one system in two successive steps; the comparison of both measurements determines the direction (and possibly the size) of the next control step. The idea of the following method is, in contrast, to measure the criterion simultaneously in two almost congruent systems differing only by a constant relative perturbation  $\Delta$  of the parameter that is to be optimized<sup>8</sup>. The difference of both measurements, related to the parameter perturbation, gives the difference quotient of the criterion

$$\frac{\bar{x_w^2}(K + \Delta \cdot K) - \bar{x_w^2}(K)}{\Delta \cdot K}$$

as an approximation to the desired differential quotient. Figure 4 shows the method applied to this example. The model (index  $M$ ) simulates the whole control system, called plant (index  $A$ ), and receives the same input signals. The mean square errors of both systems are subtracted; one common smoothing filter is sufficient for the difference  $e = u_A - u_M = x_{wA}^2(t) - x_{wM}^2(t)$  because of the linearity of the subtracting point. The total gain of the model loop is denoted by  $K_M = K$ ; the gain of the plant loop is assumed to be perturbed by  $\Delta \cdot K$ , thus  $K_A = K(1 + \Delta)$ . With eqn (4) the difference of both mean squares gives the value  $Q$  measured by the adaptive loop

$$Q = \bar{x_{wA}^2(t)} - \bar{x_{wM}^2(t)} \approx \frac{\pi}{2} k^2 \left( \kappa^2 K - \frac{1}{K} \right) \cdot \Delta \quad (13)$$

neglecting higher powers of  $\Delta$ . This equation corresponds completely to that of (10) for the periodic searching signal; it is again a measure for the differential quotient of the criterion as a function of  $K$ , having the same advantage in respect of a continuous and methodic parameter control. In addition there is an important advantage concerning the measuring time. An evaluation similar to that for the searching method yields the variance  $\sigma_q$  of the measured variable  $q$  at the low-pass filter output (see Appendix II)

$$\sigma_q \approx \frac{\pi k^2 \Delta}{8K} \cdot \frac{[(3K^2\kappa^2 - 1)^2 + 4KT(K^2\kappa^2 - 3)^2]^{\frac{1}{2}}}{KT} \quad (14)$$

the adaptive loop being opened again behind the low-pass filter. Higher powers of  $\Delta$  have been neglected, as well as again the terms small with respect to  $KT$ .

Setting for example  $\kappa \cdot K = 2$  as above, this becomes, related to the true value  $Q$  given by eqn (13)

$$\frac{\sigma_q}{Q} \approx \frac{(121 + 4KT)^{\frac{1}{2}}}{12KT} \quad (15)$$

$Q$  and  $\sigma_q$  are proportional to  $\Delta$  in a first-order approximation; consequently  $\sigma_q/Q$  and thus  $T$  do not depend on the magnitude of the perturbation. As in the case of the periodic searching signal, only the sign of  $Q$  is to be determined correctly; again it is sufficient to set  $2\sigma_q = Q$  and to have a smoothing time constant of

$$T = \frac{1.8}{K} \quad (16)$$

that is approximately 1/1,000 the value of the other two methods. This apparent difference may be explained generally by the fact that much more information is incorporated in the self-adjusting device within the model requiring a detailed knowledge of the plant. The device itself therefore needs to gather less information implying a saving in time.

The short measuring time of devices with a reference model can be utilized, not only for the adjustment to an optimum, as shown here, but, in general, for the measurement of parameters and consequently also for the self-adjustment to a predeterminate value<sup>9, 10</sup>. Another advantage is that the normal operating signals are sufficient for the self-adjustment without any need of disturbing test signals.

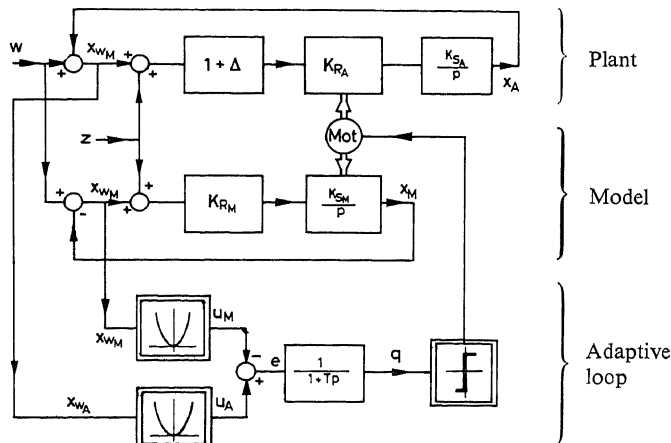


Figure 4. Plant with reference model

The identity of plant and model, excepting the one parameter, was postulated above in order to get the desired difference signal. However, inequalities of the remaining parameters, and also additional disturbing signals not considered in the model, would generate additional undesired difference signals and pretend a false optimum. This difficulty may sometimes be circumvented by tracking the plant parameters that one expects to vary, and transferring the changes to the model, using similar or other common methods. However, this causes an essential complication.

## Appendix I

In estimating the variance  $\sigma_q$  of the measured variable  $q$  greater omissions are permissible than in determining the true value  $Q$ , as shown under the heading 'Sinusoidal Searching Signal'. In the output signal  $u(t)$  of the squarer, only minor components are oscillating periodically, arising from the preceding modulation of the gain  $K' = K(1 + m_1 \cdot \sin \omega_s t)$  with  $m_1 \ll 1$ ; this periodic component can now be neglected. In computing  $Q$ , the steady component behind the phase-sensitive detection, this omission was, of course, impossible because otherwise no constant component at all would have come out. After the squarer, however, the multiplication of  $u(t)$  by  $s(t) = m_2 \cdot \sin \omega_s t$  or modulation with an index 1 is not negligible. Therefore eqn (6b) can be used for the spectral density  $S_{1uu}(\omega)$  of the fluctuating component behind the squarer with invariant  $K$ , and only the influence of the multiplication by  $s(t)$  has to be analysed.

It is practical to start from the autocorrelation function  $\phi_{vv}(\tau)$  of the multiplier output  $v(t)$  defined as an ensemble average. For a given instant  $t$

$$\phi_{vv}(\tau) \Big|_t = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} v_t \cdot v_{t+\tau} \cdot W[v_t; v_{t+\tau}] dv_t dv_{t+\tau}$$

Now  $v(t) = u(t) \cdot s(t)$ . For a certain instant  $t$  or  $(t + \tau)$  the random variable  $u(t)$  forms an ensemble; the periodic signal  $s(t)$ , in contrast, is a fixed number that can be written before the integral sign:

$$\begin{aligned} \phi_{vv}(\tau) \Big|_t &= s_t \cdot s_{t+\tau} \cdot \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} u_t \cdot u_{t+\tau} \cdot W[u_t; u_{t+\tau}] du_t du_{t+\tau} \\ &= s_t \cdot s_{t+\tau} \cdot \phi_{uu}(\tau)_t \end{aligned}$$

The time average of this is

$$\phi_{vv}(\tau) = \phi_{ss}(\tau) \cdot \phi_{uu}(\tau)$$

and with  $s = m_2 \cdot \sin \omega_s t$

$$\phi_{vv}(\tau) = \frac{m_2^2}{2} \cdot \cos \omega_s \tau \cdot \phi_{uu}(\tau)$$

The constant component in  $\phi_{uu}$ , multiplied by  $\cos \omega_s \tau$ , produces a 'deterministic' error of measurement which is not of interest here. Without it, the Fourier transform gives the spectrum of the fluctuating component of  $v(t)$

$$S_{1vv}(\omega) = \frac{m_2^2}{4} [S_{1uu}(\omega_s - \omega) + S_{1uu}(\omega_s + \omega)]$$

Inserting eqn (6b) one obtains for the example

$$S_{1_{vv}}(\omega) = \frac{m_2^2 \pi k^4 (1 + K^2 \kappa^2)^2}{2K} \left[ \frac{1}{4K^2 + (\omega_s - \omega)^2} + \frac{1}{4K^2 + (\omega_s + \omega)^2} \right]$$

The spectrum behind the low-pass filter is

$$S_{1_{qq}}(\omega) = \frac{1}{1 + \omega^2 T^2} \cdot S_{1_{vv}}(\omega)$$

and gives after integration and some algebraic manipulations, setting again  $KT \gg 1$ ,

$$\sigma_q^2 = \int_0^\infty S_{1_{qq}}(\omega) d\omega = \frac{\pi^2 m_2^2 k^4 (1 + K^2 \kappa^2)^2}{4K^2} \cdot \frac{2KT}{T^2(\omega_s^2 + 4K^2)}$$

Earlier in the text a searching signal slow with respect to the cut-off frequency of the control loop was assumed, i.e.  $\omega_s \ll K$ . Hence the ratio of the variance to the desired value  $Q$  given by eqn (10) becomes

$$\frac{\sigma_q}{Q} \approx \frac{\sqrt{2} \cdot (K^2 \kappa^2 + 1)}{(K^2 \kappa^2 - 1) \cdot m_1} \cdot \frac{1}{(KT)^{\frac{1}{2}}} \quad (11a)$$

## Appendix II

The spectra of the a.c. components behind the squarer become, for the plant and model of Figure 4 and using eqn (6b),

$$S_{1_{u_A u_A}}(\omega) = \frac{2\pi k^4 [1 + \kappa^2 \cdot K^2 (1 + \Delta)^2]^2}{K(1 + \Delta)} \cdot \frac{1}{4K^2 (1 + \Delta)^2 + \omega^2} \quad (17)$$

and

$$S_{1_{u_M u_M}}(\omega) = \frac{2\pi k^4 [1 + \kappa^2 K^2]^2}{K} \cdot \frac{1}{4K^2 + \omega^2} \quad (18)$$

respectively.

For the spectrum  $S_{1_{ee}}(\omega)$  of the difference  $e = u_A - u_M$  the cross power spectra must be considered because of the statistical correlation of  $u_A$  and  $u_M$ :

$$\begin{aligned} S_{ee}(\omega) &= S_{u_M u_M} + S_{u_A u_A} - S_{u_M u_A} - S_{u_A u_M} \\ &= S_{u_M u_M} + S_{u_A u_A} - 2 \operatorname{Re} \{S_{u_M u_A}\} \end{aligned} \quad (19)$$

where

$$\boxed{\operatorname{Eqn} (20)}^*$$

Eqns (17), (18) and (20) for the spectra are inserted into (19).

After some algebraic manipulations the terms with  $\Delta^0$  and with  $\Delta^1$  cancel, and omitting  $\Delta^3$  and higher powers of  $\Delta$

$$S_{1_{ee}}(\omega) \approx \frac{\pi k^4 \Delta^2}{2K} \cdot \frac{4K^2 (K^2 \kappa^2 - 3)^2 + \omega^2 (3K^2 \kappa^2 - 1)^2}{(4K^2 + \omega^2)^2} \quad (21)$$

The spectrum behind the low-pass filter

$$S_{1_{qq}}(\omega) = \frac{1}{1 + T^2 \omega^2} \cdot S_{1_{ee}}(\omega)$$

is integrated, giving

$$\sigma_q \approx \frac{\pi k^2 \Delta}{8K} \cdot \frac{[3K^2 \kappa^2 - 1]^2 + 4KT(K^2 \kappa^2 - 3)^2]^{\frac{1}{2}}}{KT} \quad (14)$$

with  $KT \gg 1$ .

## References

- 1 DAVENPORT, W. B., JOHNSON, R. A., and MIDDLETON, D. Statistical errors in measurements on random time functions. *J. Appl. Phys.* 23 (1952), 377
- 2 DOGANOVSKII, S. A. Automatic operating mode optimization of a class of systems with respect to statistical performance criteria. *Automat. Telemekh.* 21 (1960), 1105; English transl. *Automation and Remote Control, Instrum. Soc. Amer.* 779, 1961
- 3 BENNETT, R. R., and FULTON, A. S. Measurement of low frequency random noise. *J. Appl. Phys.* 22 (1951), 1187
- 4 LANING, J. H., and BATTIN, R. H. *Random Processes in Automatic Control*. 1956. New York; McGraw-Hill
- 5 SCHLITT, H. *Systemtheorie für regellose Vorgänge*. 1960. Springer-Verlag
- 6 MCGRATH, R. J., and RIDEOUT, V. C. A simulator study of a two-parameter adaptive system. *IRE-Trans. Autom. Contr.* AC-6 (1961), 35
- 7 KAZAKOV, I. E. The dynamics of self-adaptive systems with extremal continuous adjustment of the compensating network in the presence of random disturbances. *Autom. Telemekh.* 21 (1960), 1465-1474. English transl. *Automation and Remote Control, Instrum. Soc. Amer.* 1040, 1961
- 8 BRAMMER, K. Selbsteinstellender Regler mit selbsteinstellendem Bezugsmodell. Studienarbeit. *Inst. Regelungstechnik*, Darmstadt, (1961)
- 9 SCHMIDT, G. Selbsteinstellender Regelkreis mit Bezugsmodell (Auszug einer Studienarbeit, *Inst. Regelungstechnik* Darmstadt, 1960); *Regelungstechnik* 10 (1962), 145
- 10 MESCH, F. Selbsteinstellender Regelkreis mit selbsteinstellendem Bezugsmodell. *Zs. f. Messen, Steuern, Regeln* 5 (1962), 320

\* Eqn (20):

$$\operatorname{Re} \{S_{u_M u_A}\} = \frac{2\pi k^4 (1 + \kappa^2 K^2) \cdot [1 + \kappa^2 K^2 (1 + \Delta)^2] \cdot [4K^2 (1 + \Delta) + \omega^2]}{K \cdot (1 + \Delta)^{\frac{1}{2}} \cdot (4K^2 + \omega^2) \cdot [4K^2 (1 + \Delta)^2 + \omega^2]} \quad (20)$$

## DISCUSSION

B. QVARNSTRÖM, *Chalmers Institute of Technology, Gibraltargatan 5 R, Gothenburg S, Sweden*

In his very interesting paper Mr. Mesch considers an adaptive control loop, in which a gain parameter is adjusted by use of a measured quantity  $q(t)$ . This quantity yields in some way or other the information about the performance of the main control loop judged on the basis of a quadratic criterion. Precisely how the gain parameter will be controlled by the quantity  $q(t)$  is not discussed by the author, but there is no doubt that an adaptive control action can be realized.

The stepwise searching method suggested incorporates an adaptive control of the gain parameter, designed to minimize the mean square error  $Q = \bar{x_w^2}$ , or, when  $Q$  is unavailable, to minimize the quantity  $q(t)$ , which is the best estimation of  $Q$  under certain conditions. The author has shown in eqns (4) and (5) that  $Q$  will reach a minimum for a certain value of the gain parameter  $K$ , denoted by  $K_{opt}$ . This value of the gain factor is considered, by definition, to be the best or optimal value of the gain.

The choice of an optimizing criterion is an important one and I

would like to know more in detail how the author will justify the choice in this case, because there is a slight contradiction in the paper, as I see it. On one hand the minimum of the mean square error is expressed by eqn (4) under the assumption that the gain parameter maintains a constant value during the whole averaging or smoothing period. On the other hand, the task of the adaptive loop is to continuously control the gain parameter, which means that the gain might vary or fluctuate if a non-constant gain is preferable; and, this is the point, a suitable gain parameter variation may in this case reduce the mean square error under the minimum value given by a constant gain.

The measuring time calculated by the author is a consequence of a restriction upon the fluctuation of the measured quantity  $q(t)$ . From a practical point of view, I think this restriction might be necessary, but I can hardly see any fundamental reason why the fluctuations of  $q(t)$  should be kept within these or those limits. There is no harm in a fluctuating signal, especially not, as I see it, when the variation yields information about the signals fed into the system, an information that can be used to reduce the mean square error. It is easier for me to accept a restriction on the fluctuations of the gain parameter  $K$ , because there must be a number of practical reasons not to allow a wild fluctuation of the gain factor in a real plant.

There is a deep similarity between the stepwise and the sinusoidal searching methods. One would expect a smoothing time of the same order of magnitude in both cases, as the author already has shown in his paper. Now, the function  $Q$  is not quite the same in these two examples, and I am not sure that the restrictions used therefore are equivalent. I would appreciate the author's comments on this point, because I found it interesting that an intentional perturbation did not promote the adaptive control possibilities.

In the third example, the model reference system, the author demonstrates a remarkable scheme permitting the quantity  $Q$  to be measured 1,000 times faster than before. By the use of a model, almost identical with the plant, a major part of the statistical fluctuations of the squared error is cancelled. In the beginning of the paper the author states that the gain of the process  $K_S$  may change in an unpredictable manner. I cannot see that the model reference system would work properly under this condition, because a specific and known gain difference between model and plant should be necessary.

The whole system, plant and model, forms a kind of differentiating device. It is well known that devices of that kind are very sensitive to disturbances. One consequence of a disturbance other than  $z$ , that is measured and also fed into the model, must be a contribution to the fluctuations of the controlled gain parameter  $K$ . I feel that in neglecting this source of disturbance the author might have overestimated the performance of the model reference system.

Finally, I would like to say that I agree with Mr. Mesch in most of his conclusions and I have found his work stimulating and important.

#### F. MESCH, in reply

Answering Professor Qvarnström, a quickly-varying gain  $K$  would mean an essentially and intentionally non-linear system; for Gaussian input signals as assumed in the paper, however, it is well known that linear filters are optimal in the Wiener sense. Thus linear basic loops were considered with slowly-varying parameters. The second comment is quite true; in the first case, the function  $Q$  is the error criterion to be optimized, and in the two other cases it is the derivative of this criterion with respect to the parameter  $K$ . I am in complete agreement with the third comment. As already pointed out in the oral presentation of the paper, the model reference system is sensitive to noise signals not fed simultaneously into the model. Therefore the model system might be used only as a double model, e.g. on an analogue computer. But recently we developed another configuration combining the model system with the sinusoidal perturbation system. Preliminary results indicate that this more elaborate configuration has a short measuring time due to the model, but is also insensitive to non-considered noise, due to correlating the output signal with the test signal.

D. ŠILJAK, *Electrotechnical Faculty, University of Belgrade, Macvanska 8, Belgrade, Yugoslavia*

The comparison of the measuring times of controller gain in various first-order self-adjusting systems is straightforward, the results indicating that the measuring time of the gain to be adjusted is of an essential interest. However, the proposed approach, if applied to higher-order systems, may suffer from the limitation that the existence of a finite optimum value of adjustable gain is assumed. As known<sup>1</sup>, if a control system of the second, or higher order, is optimized according to the mean-squared error as a performance index, it may happen that only the infinite value of the system gain makes the optimum value of the chosen index. In order to obtain physically realizable and practically applicable values of adjustable system parameters, it is necessary to optimize the mentioned performance index subject to the constraints<sup>1-4</sup>. The introduction of constraints may change significantly the system dynamics, thus affecting the measuring time. It may, therefore, be suggested that the author introduce certain constraints into the optimization of the controller gain, and make his method applicable to more general system configurations.

#### References

- 1 NEWTON, G. C., JR., GOULD, L. A. and KAISER, J. F. *Analytical Design of Linear Feedback Controls*. 1957. New York; Wiley
- 2 GAYLORD, R. Dual input systems with a saturation constraint. *2nd I.F.A.C. Congr.* Basle, 1963. London; Butterworths: Munich; Oldenbourg
- 3 ŠILJAK, D. Optimization of squared error with a relative stability constraint. *Ph. D. Thesis*. University of Belgrade, March 1963
- 4 ŠILJAK, D. Generalization of Mitrovic's method, *I.E.E.E. Summer General Meeting*, Paper No. 63-988, 1963. Toronto

#### F. MESCH, in reply

I agree with Dr. Šiljak that an optimum must exist if it is to be found by the self-adjustment, and that constraints might be important and might influence the form of the basic loop. However, the basic loop is the same in all three cases; moreover, recent results indicate that the model system is quite generally superior to the other two systems, independent of the basic loop considered.

F. B. TUTEUR, *Department of Engineering & Applied Science, Yale University, Conn., U.S.A.*

The startlingly large improvement in measuring time obtained by the author with his third system relative to the first two systems appears to be due entirely to the fact that the model and system receive precisely the same signal and noise. The measurement is therefore essentially a noiseless measurement. This is not true of the first two methods considered.

It would also be interesting to have the author's comment on whether there is a physical reason why eqn (15) gives a linear short-time relation between R.M.S. error and estimation time, whereas the more usual relation is that the R.M.S. error varies inversely as the square root of the time [as in eqns (7) and (11b) and, for instance, as given in the paper by B. Qvarnström].

#### F. MESCH, in reply

I disagree with Professor Tuteur's first comment. The model system is superior to the other two systems even if there is only one signal in all cases, without noise. The reason is, as indicated in the paper, that in the model there is more *a priori* information incorporated resulting in a shorter time for gathering the rest of information required. The simple square-root relation between variance and measuring time occurs, generally, only for very long measuring time constants. For shorter ones, as obtained with the model, the relationship is more involved, as might be seen, e.g., from References 1 and 3 of the paper.

O. L. R. JACOBS, *Department of Electrical Engineering, Edinburgh University, Edinburgh 9, Scotland*

My question concerns the third system described in the paper; this system uses a model with parameter perturbation and is shown in Figure 4 of the paper. It can be seen that the random disturbance  $Z$ , which is partly responsible for errors in the original system, is assumed to act on the model as well as on the plant. This implies that disturbance  $Z$  can be measured and made available to a controller.

If it is possible to measure  $Z$ , and if the plant is sufficiently well understood that it can be modelled as proposed, would it not be preferable to control the variable gain  $K$  by some direct feedforward control from the measured value of  $Z$ ?

F. MESCH, *in reply*

Dr. Jacobs' proposal seems to me impractical. It is not sufficient to measure only the disturbance  $Z$ ; it would be necessary also to measure its statistics (power spectrum) and to automatically compute the optimum parameter setting. This would be much more complicated.

G. ROSENAU, *Deutsche Forschungsanstalt für Luft- und Raumfahrt, Steinhorstwiese 20a, Braunschweig, Germany*

(1) Mr. Mesch gives the power spectral densities for the input signal  $S_{ww} = k^2/w^2$  and for the disturbance  $S_{zz} = K^2 k^2$ .

(a) What were the reasons for choosing these power spectra?

(b) Did the author investigate other power spectra which would be more realistic for applications as, for example, in aeronautics?

(c) Did the author attempt to generalize his results to the case where the input signal  $w$  and the disturbance  $z$  have different power spectra, e.g. having different frequency bands.

(2) Is it possible to generalize the model reference systems to the case where one does not know the disturbance, which will be the general case, and where, therefore, it is impossible to introduce the disturbance into the model (see Figure 4).

F. MESCH, *in reply*

I disagree with Mr. Rosenau. I would like to know what kind of spectra would seem to him more realistic. Command input signals

usually have spectra decreasing with frequency, and disturbances are usually broadband. As the spectra are the same in all three cases, the comparison is little affected by their specific form which has been chosen, therefore, for computational ease. The second question I have already answered in my reply to Professor Qvarnström.

C. M. WOODSIDE, *Cambridge University, Cambridge, England*

I wish to suggest an alternative formulation of Mr. Mesch's model approach, which demonstrates its similarity to the step search method.

He postulates that the random input and noise are available to the model. Therefore a tape recording of them can be made. If the search proceeds, as in the stepping method with plant alone, and two estimates of mean squared error are made in two consecutive 'runs' with  $K_R$  perturbed by a factor  $1 + \Delta$  in the second run, then if the tape recording is started in the same place at the beginning of each run, the first run corresponds with the model in the paper, and the second with the plant. The difference between estimates corresponds exactly with  $Q$  in eqn (15).

The vast part of the time reduction in obtaining  $Q$ , which is still approximately  $1/500$ , is then seen to be due to knowing the input and disturbance exactly rather than knowing the plant exactly. Knowing the plant reduces the time by a further half.

In comparing sinusoidal and stepping search, it would be more instructive to compare the derivative signal in the former with the difference between two successive estimates of mean square error in the latter, rather than with one estimate of mean square error as in the paper. The times required to give 5 per cent accuracy are then equal.

F. MESCH, *in reply*

I find Mr. Woodside's proposal very interesting because it eliminates the need of a model, and yet yields a short measuring time. However, it assumes that arbitrary (tape recorded) input signals can be applied to the system. On the contrary, in the paper it was assumed that only normal operating input signals are allowed. I agree, too, that it is more correct to consider the *difference* of two measurements, and indeed we have been doing this for some time.

# On Self-adjusting Control Systems Without Test Disturbance Signals

E.P. POPOV, G.M. LOSKUTOV and R.M. YUSUPOV

## Summary

The report presents and analyses one of the principles for designing an adaptive system to control time-varying processes. According to this principle, in the processes of self-adjusting, the dynamic characteristics of the system (plant and controller) and the effect of some uncontrolled external disturbances are taken into consideration. This controlled process is close, on the whole, to the predetermined (reference) behaviour in adjusting controller parameters. A high-speed computer, either a digital or an electronic analogue one, is supposed to be included in the loop of the adaptive control, depending upon the method of the process identification. Dynamic characteristics of the system are found by differential equations with time-constant coefficients, by means of which real and reference processes are approximated.

In estimating reference values of controller coefficients, the sum of the squares of real coefficient value deviations from the reference ones is minimized. Mechanization problems of computing the above coefficients by definite servos are discussed. The report deals with some particular applications of the given method for designing such an adaptive control system.

## Sommaire

Ce rapport met en évidence et étudie l'un des principes permettant d'élaborer un système adaptatif destiné à commander des processus évoluant dans le temps. Conformément à ce principe, on prend en considération, pour l'auto-adaptation les caractéristiques dynamiques du système (installation et commande) et l'effet de quelques perturbations externes incontrôlées. Le processus commandé est, d'une manière globale, rendu proche du comportement prédéterminé (référence) par réglage des paramètres de la commande. On suppose qu'une calculatrice rapide du type numérique ou analogique, est insérée dans la boucle de la commande adaptative, compte tenu de la méthode d'identification utilisée. Les caractéristiques dynamiques du système sont données par des équations différentielles à coefficients indépendants du temps, équations qui permettent d'approximer les processus réels comme les processus de référence. Lors de l'estimation des valeurs de référence des coefficients du régulateur, on minimise la somme des carrés des écarts des coefficients réels par rapport aux coefficients de référence. Les problèmes liés à la mécanisation du calcul de ces coefficients à l'aide de servo-mécanismes sont passés en revue. Le rapport s'achève par la description de quelques applications particulières de la méthode proposée à l'élaboration de commandes adaptatives de ce type.

## Zusammenfassung

Dieser Beitrag behandelt ein Bauprinzip für anpassende Regelung zeitveränderlicher Prozesse. Hierbei werden für die Selbstanpassung die dynamischen Eigenschaften des Systems (Strecke und Regler) und die Auswirkungen einiger ungeregelter äußerer Störeinflüsse berücksichtigt. Bei der Einstellung der Regelparameter entspricht dieser Vorgang im großen ganzen dem Sollverhalten. Je nach der Art der Regelstrecke ist der Schnellrechner im selbstanpassenden Regelkreis entweder digital oder elektronisch analog. Die dynamischen Eigenschaften des Systems ergeben sich aus Differentialgleichungen mit Zeitkonstanten-Koeffizienten, durch welche die wirklichen den bezogenen Prozessen angenähert werden.

Die Summe der quadrierten Abweichungen dieser beiden Prozesse wird zur Vorausschätzung der Bezugswerte der Reglereinstellungen

minimiert. Es wird untersucht, wie diese Koeffizienten durch vorbestimmte Steuerungssysteme mechanisiert werden können. Der Beitrag behandelt einige Anwendungen dieser Methode für den Entwurf eines solchen anpassenden Regelsystems.

## Statement of the Problem

In this paper, the term 'self-adjusting control system' means a system which performs the following three operations:

(1) Measures by means of automatic search or computes from the results of measurements the dynamic characteristics of the system, and possibly the characteristics of the disturbances as well.

(2) On the basis of this or that criterion defines the controller setting, parameters or structure needed for calibration (or optimization).

(3) Realizes the resultant controller structure, parameter or setting values.

Many studies of the theory and practice of self-adjusting control systems for stationary controlled plants have so far appeared in the world literature. There have also been contributions on self-adjusting of quasi-stationary systems. But there is almost a complete lack of contributions dealing more or less specifically with problems of synthesis and analysis of self-adjusting control systems for essentially non-stationary controlled plants. Moreover, as far as the authors are aware, even in the case of stationary and quasi-stationary systems, the process of self-adjustment is frequently effected solely on the basis of an analysis of the dynamic characteristics of the system, without taking into account the unmeasured external disturbances acting upon the controlled plant. At the same time it is obvious that external disturbance, besides the dynamic characteristics of the system, determines the quality of the control process.

Another drawback of many of the self-adjusting systems in existence and proposed in the literature is the need to use special test signals to check the dynamic characteristics of the system.

This paper proposes, and attempts to validate, one of the possible principles for the creation of a self-adjusting control system for a particular class of non-stationary controlled plants.

The main advantage of the principle in question is the opportunity it provides to take account of both internal (system parameters) and external (harmful and controlling disturbances) conditions of operation of the system. In contrast to the self-adjusting systems known, a system created in accordance with the principle proposed will make it possible to obtain automatically the fullest possible information about the process under control without the use of test signals.

For the operation of a self-adjusting control system created on the basis of the principle proposed, a mathematical model of a reference (calculated) control system must be constructed. A 'reference system' is understood to be a system the controller of which is designed in accordance with the requirements on the quality of the control process, with the assumption that the mode of variation in time of the system's parameters as well as the disturbance effects are known.

The structure of the mathematical approximation of the real process is selected to match that of the mathematical model of the reference process. The self-adjusting system operates in such a way as to ensure continuous identity between the mathematical approximation of the real process and the model of the reference system. In this connection, the problem is posed of making the mathematical approximation of the real process as close as possible to the model of the reference process.

Without loss of generality, the case of control of only one variable is considered, which is denoted by  $x$ , and the corresponding reference differential equation is written in the form

$$x_E^{(n)} + \sum_{i=0}^{n-1} a_i^E(t) x_E^{(i)} = \sum_{i=0}^m b_i^E(t) f_E^{(i)} \quad (1)$$

The real process is approximated by a linear differential equation of the same structure:

$$x^{(n)} + \sum_{i=0}^{n-1} a_i(t) x^{(i)} = \sum_{i=0}^m b_i(t) f^{(i)} \quad (2)$$

$$t = t_0, x^{(i)}(t_0) = x_{E0}^{(i)} \quad (i=0, 1, \dots, n-1)$$

The operation of the proposed self-adjusting control system will be examined in accordance with the sequence of the process of self-adjustment, indicated at the beginning of the definition.

### General Case of Determination of the Dynamic Characteristics of a System

In order to create an engineering method of determining the dynamic characteristics of non-stationary systems in the construction of a self-adjusting control system, this paper proposes the use of the methods of stationary systems. For this purpose, the non-stationary system (1) is replaced by an equivalent system with piecewise-constant coefficients. (The methods of stationary systems are used in the intervals of constancy of the coefficients.) The transfer from a system with variable coefficients to one with piecewise-constant coefficients is effected on the basis of a theorem which can be formulated with the assistance of a number of the propositions of the theory of ordinary differential equations. In accordance with this theorem, the solution of a differential equation of form (1) with piecewise-continuous coefficients (a finite number of discontinuities of the first kind is assumed) can be obtained with any degree of accuracy in a preset finite interval  $(t_0, T_0)$  by breaking down the latter into a finite number of sub-intervals  $(t_K, t_{K+1})$  and replacement of the variable coefficients within each sub-interval by constants, equal to any values of the corresponding coefficients inside or on the boundaries of the sub-intervals under consideration. In the general case, it is expedient to effect the breakdown process by the method of multiple iteration of solutions on a high-speed computer.

Let the differential equation with variable coefficients (1) be approximated by an equation with piecewise-constant coefficients.

Then, for  $t \in (t_K, t_{K+1})$ , one may write

$$x_E^{(n)} + \sum_{i=0}^{n-1} a_{iK}^E x_E^{(i)} = \sum_{i=0}^m b_{iK}^E f^{(i)} \quad (3)$$

In accordance with differential equation (3), the real process is approximated by the equation

$$x^{(n)} + \sum_{i=0}^{n-1} a_{iK} x^{(i)} = \sum_{i=0}^m b_{iK} f^{(i)} \quad (4)$$

As dynamic characteristics of the system at the first stage of operation of the self-adjusting system in each interval  $(t_K, t_{K+1})$ , the coefficients  $a_{iK}$  ( $i = 0, 1, \dots, n-1$ ),  $b_{iK}$  ( $i = 0, 1, \dots, m$ ) are defined.

The simplest way to define these coefficients lies in defining the values of  $x$  and  $f$  and their corresponding derivatives at the points  $t_K = \tau_1, \tau_2, \dots, \tau_s = t_{K+1} - \Delta t$ .

By substituting these values into eqn (4), one obtains for each interval  $(t_K, t_{K+1})$  a system of  $S$  algebraic dissimilar equations for defining the searched coefficients.

In practice it is not always possible to measure the disturbing effect  $f$  and its derivatives. Therefore, in the general case, the above-mentioned method of defining the coefficients  $a_{iK}$  and  $b_{iK}$  cannot be directly employed.

This difficulty may be avoided in the following way. The real process is approximated, not by differential eqn (4), but by a differential equation of the form

$$\bar{x}^{(n)} + \sum_{i=0}^{n-1} \bar{a}_{iK} \bar{x}^{(i)} = \sum_{i=0}^m \bar{b}_{iK} f_E^{(i)} \quad (5)$$

In eqn (5) the disturbing effect and its corresponding derivatives are taken to equal the reference values. This avoids the need to measure the real disturbance  $f$ , and makes it possible to use the above-mentioned means of defining the coefficients of the differential equation approximating the real control process. The non-agreement of the real disturbances with the reference ones are taken into account through the coefficients  $a_{iK}$  and  $b_{iK}$ . Therefore dashes are placed over them.

In the general case  $\bar{x}^{(i)} \neq x^{(i)}$  ( $i = 0, 1, \dots, n$ ) i.e., there is an approximation error. In view of this, in the transfer from eqn (4) to eqn (5), it is necessary to evaluate the maximum possible value of this approximation error, using for this purpose the assumed values of the limits of variation of disturbance  $f$ .

If for some class of controlled plants it can be assumed that in the process of operation only the scale of the disturbance changes, i.e., the equality

$$f(t) = C_K f_E(t), \quad t \in (t_K, t_{K+1}) \quad (6)$$

where  $C_K$  is the random scale of disturbance, is satisfied, then the approximation error is absent, and the connection of the coefficients of eqns (4) and (5) is expressed by the equalities:

$$\begin{aligned} \bar{a}_{iK} &= a_{iK} \quad (i=0, 1, \dots, n-1) \\ \bar{b}_{iK} &= C_K b_{iK} \quad (i=0, 1, \dots, m) \end{aligned} \quad (7)$$

Equation (5) is used (henceforward, to simplify the notation, the dashes over the coefficients and the variable  $x$  are dropped)



for definition of the coefficients  $a_{iK}$  and  $b_{iK}$ . It is assumed that measurements  $x, x', \dots, x^{(n)}$  are performed at the points  $t_K = \tau_1, \tau_2, \dots, \tau_S = t_{K+1} - \Delta t$ .

The values of  $f_E, f'_E, \dots, f_E^{(m)}$  are known. Then, for the definition of  $(n + m + 1)$  desired coefficients in each interval  $(t_K, t_{K+1})$  one obtains the following system of  $S$  algebraic equations, which will be written in abbreviated form thus:

$$\sum_{i=0}^{n-1} x^{(i)}(\tau_j) a_{iK} - \sum_{i=0}^m f_E^{(i)}(\tau_j) b_{iK} = -x^{(n)}(\tau_j) \quad (j=1, 2, \dots, S) \quad (8)$$

It is not always expedient to solve directly system (8) for  $S = m + n + 1$ , since, on account of the existence of measuring instrument errors and random high-frequency control process oscillations, the accuracy of definition of the coefficients will be very low. Moreover, for the same reasons, system (8) may be altogether incompatible.

To eliminate the case of incompatibility and to increase the accuracy of definition of the searched coefficients the method of least squares is employed<sup>1, 2</sup>. In so doing, the problem of approximation is also solved. When utilizing this method, it is expedient to take  $S > m + n + 1$ .

Using the method of least squares, the coefficients  $a_{iK}, b_{iK}$  are defined, minimizing according to these coefficients the function

$$L = \sum_{j=1}^S \rho(\tau_j) L_j^2$$

where

$$L_j = \sum_{i=0}^{n-1} x^{(i)}(\tau_j) a_{iK} - \sum_{i=0}^m f_E^{(i)}(\tau_j) b_{iK} + x^{(n)}(\tau_j)$$

is the disagreement, and  $\rho(\tau_j)$  are weight coefficients which define the value of each measurement and, accordingly, of each of the equations of system (8).

The necessary condition of the minimum of function  $L$  is the equality to zero of its first-order partial derivatives according to  $a_{iK}$  and  $b_{iK}$ . Having computed the partial derivatives and equated them to zero, one obtains an already compatible system of  $m + n + 1$  linear algebraic equations for the definition of  $m + n + 1$  coefficients:

$$\begin{aligned} \frac{\partial L}{\partial a_{iK}} &= \sum_{j=1}^S \rho(\tau_j) L_j \frac{\partial L_j}{\partial a_{iK}} = 0 \quad (i=0, 1, \dots, n-1) \\ \frac{\partial L}{\partial b_{iK}} &= \sum_{j=1}^S \rho(\tau_j) L_j \frac{\partial L_j}{\partial b_{iK}} = 0 \quad (i=0, 1, \dots, m) \end{aligned} \quad (9)$$

Solving system (9) by known methods, one obtains the values of  $a_{iK}$  and  $b_{iK}$ .

In certain cases the process of control at intervals may be approximated by a differential equation of the form

$$x^{(n)} + \sum_{i=0}^{n-1} a_{iK} x^{(i)} = \varphi_{EK}(t) \quad (10)$$

where

$$\varphi_{EK}(t) = \sum_{i=0}^m b_{iK} f_E^{(i)}(t)$$

This coarser approximation will make it possible to reduce computing time considerably by a reduction of the quantity of searched coefficients; in the given case only the coefficients  $a_{iK}$  are desired.

In the given approximation the deviations of the values of

real coefficients  $b_{iK}$  and real disturbances  $f$  will be taken into account in the system *via* the values of the coefficients  $a_{iK}$ . System (11) will be the initial algebraic system for definition of the coefficients  $a_{iK}$ :

$$\sum_{i=0}^{n-1} x^{(i)}(\tau_j) a_{iK} = \varphi_{EK}(\tau_j) - x^{(n)}(\tau_j) \quad (j=1, 2, \dots, S) \quad (11)$$

For definition of the searched coefficients  $a_{iK}$  by the method of least squares, one minimizes the function

$$L_1 = \sum_{j=1}^S \rho(\tau_j) L_j^2 \quad (12)$$

where

$$L_j = \sum_{i=0}^{n-1} x^{(i)}(\tau_j) a_{iK} + x^{(n)}(\tau_j) - \varphi_{EK}(\tau_j)$$

Using the necessary condition of the existence of a minimum of function (12) for the definition of  $n$ , coefficients  $a_{iK}$  ( $i = 0, 1, \dots, n-1$ ), one obtains a system of  $n$  algebraic equations:

$$\frac{\partial L_1}{\partial a_{iK}} = \sum_{j=1}^S \rho(\tau_j) L_j \frac{\partial L_j}{\partial a_{iK}} = 0 \quad (i=0, 1, \dots, n-1) \quad (13)$$

All the above discussion and the operations were performed on the assumption that the values of the control variable and the necessary quantity of derivatives at the moments of time of interest are available. In practice, however, one is usually limited to second-order derivatives.

In a number of cases real high-order systems may be approximated by second-order differential equations, preserving the description of their main dynamic properties. But even in the case of more complex high-order systems it is possible to suggest a number of algorithms for defining the searched coefficients, given the existence of a limited quantity of derivatives, some of which are as follows:

(a) Derivatives of higher orders of the control variable can be calculated with the assistance of a digital computer on the basis of the Lagrange and Newton interpolation formulae or according to the formulae of quadratic interpolation (method of least squares).

(b) If one integrates each term of eqns (5) and (10)  $n - q$  times, where  $q$  is the order of the senior derivative of the control variable, which one can measure in a system with the requisite accuracy, then, taking the limits of integration  $t_K, \tau_j$  ( $j = 1, 2, \dots, S$ ), one obtains the integral forms of eqns (8) and (11) respectively. If reference values are given to the magnitudes  $x^{(n-1)}(t_K), x^{(n-2)}(t_K), \dots, x^{(n-q+1)}(t_K)$  in these equations, then for defining the coefficients  $a_{iK}$  ( $i = 0, 1, \dots, n-1$ ) and  $b_{iK}$  ( $i = 0, 1, \dots, m$ ) it is sufficient to measure the derivatives to the  $q$ th order.

(c) Practically all existing controlled plants and control systems can be described by a set of differential equations, each of which characterizes one degree of freedom of movement and therefore has an order no higher than second.

(d) Sometimes, to reduce the order of the derivatives required for measurement, one may also take advantage of a number of coarse assumptions in relation to the terms of eqns (5) and (10), which contain derivatives of high orders.

For example, in these equations the values of the derivatives  $x^{(n)}, x^{(n-1)}, x^{(n-q+1)}$  can be assumed equal to the reference values.



(e) The coefficients of approximating eqns (5) and (10) can be defined without any recourse to algebraic systems (8) and (11), if one uses the following method<sup>5</sup>.

Let the composition of the control system include an analogue simulator, on which is set up a differential equation of form (5) or (10). In this simulator there is a controlling device, which provides an opportunity to effect variation of coefficients  $a_{iK}$  and  $b_{iK}$  in a certain way.

The control system memorizes the curve of the real process in the interval  $(t_K, t_{K+1} - \Delta t)$ , and selection of the coefficients  $a_{iK}$  and  $b_{iK}$  is performed on the simulator in such a way as to bring together in a certain sense the real process and the solution of the equation set up on the simulator.

When the quantitative value of the proximity evaluation reaches the predetermined value, the magnitudes of coefficients  $a_{iK}$  and  $b_{iK}$  are fixed and extracted for subsequent employment in the self-adjusting control system. Obviously the simulator operation time scale must be many times less than the real time scale of the system. Only under this condition can the requisite high speed of self-adjustment be achieved. Practically any time scale may be realized with the assistance of analogue computing techniques.

#### Automatic Synthesis of Controller Parameters

For the operation of the majority of self-adjusting systems, the system operation quality criterion is set in advance. For systems constructed on the basis of the proposed principle, it is generally expedient to use as the criterion the expression

$$M = \sum_{i=0}^{n-1} (a_{iK} - a_{iK}^E)^2 + \sum_{i=0}^m (b_{iK} - b_{iK}^E)^2 \quad (14)$$

This criterion generalizes both methods of approximation of the real control process expounded above.

To simplify subsequent operations, the following notations are introduced.

$$b_{0K} = a_{nK}; \quad b_{1K} = a_{n+1,K}, \dots, b_{mK} = a_{m+n,K}$$

Expression (14) can then be rewritten in the form

$$M = \sum_{i=0}^{n_0} (a_{iK} - a_{iK}^E)^2; \quad n_0 = \begin{cases} n+m & \text{for (5)} \\ n-1 & \text{for (10)} \end{cases} \quad (15)$$

In each interval  $(t_K, t_{K+1})$  the adjustable parameters are so selected as to bring expression (15) to the minimum. The ideal, i.e., most favourable, case would be one when  $M$  would reach zero as the result of selection of the adjustable parameters. This is not always possible, however. In the first place, not all the coefficients  $a_{iK}$  ( $i = 0, 1, \dots, n_0$ ) are controllable. Second, in multi-loop non-autonomous systems even the values of the controllable coefficients cannot all be tuned up to the reference values simultaneously, since the relationship of the coefficients  $a_i$  to the adjustable parameters, although usually linear, is nevertheless arbitrary with respect to the quantity of adjustable parameters, the sign and the coefficients with which these parameters enter into expressions for  $a_i$ .

The second difficulty may be avoided by means of successful selection of the reference system or by complete disconnection of the loops (channels) of control from the main variables, i.e., by satisfying the conditions of autonomy.

It is assumed that all the coefficients  $a_i$  ( $i = 0, 1, \dots, n_0$ ) are controllable (in practice the values of uncontrollable coefficients may be reckoned to be reference values). Then, for the coefficients  $a_i$  one may write

$$a_i = a_i(K_1, K_2, \dots, K_p; T_1, T_2, \dots, T_q; l_1, l_2, \dots, l_r) \quad (i = 0, 1, \dots, n_0)$$

where  $K_1, K_2, \dots, K_p$  are the gains of the controlled plant;  $T_1, T_2, \dots, T_q$  are the time constants of the controlled plant and the controller, and  $l_1, l_2, \dots, l_r$  are the gains of the controller (adjustable parameters).

Since the coefficients  $a_i$  usually depend on the adjustable parameters linearly, one may write

$$a_i = \sum_{j=1}^r \mu_{ij} l_j + v_i \quad (i = 0, 1, \dots, n_0) \quad (16)$$

where

$$\mu_{ij} = \mu_{ij}(K_1, K_2, \dots, K_p; T_1, T_2, \dots, T_q);$$

$$v_i = v_i(K_1, \dots, K_p; T_1, T_2, \dots, T_q)$$

Using the necessary condition for the existence of a minimum of function  $M$ , one obtains the following algebraic system for determination of the setting values  $l_1, l_2, \dots, l_r$

$$\sum_{i=0}^{n_0} [a_{iK}(l_1, l_2, \dots, l_r) - a_{iK}^E] \frac{\partial a_{iK}(l_1, l_2, \dots, l_r)}{\partial l_j} = 0 \quad (j = 1, 2, \dots, r) \quad (17)$$

It is assumed that when the system is in operation, the adjustable parameter values only change in accordance with their computed values, i.e., at any moment of time one knows the magnitudes of  $l_1, l_2, \dots, l_r$ . Then, for the interval  $(t_K, t_{K-1})$  until the moment of correction of the adjustable parameters in accordance with expression (16), one can write:

$$a_{iK} = \sum_{j=1}^r \mu_{iK} l_{j,K-1} + v_{iK} \quad (18)$$

From system (18) one may determine the magnitudes of  $\mu_{iK}$  and  $v_{iK}$  ( $i = 0, 1, \dots, n_0$ ;  $j = 1, 2, \dots, r$ ) since the values of  $a_{iK}$  ( $i = 0, 1, \dots, n_0$ ) and  $l_{j,K-1}$  ( $j = 1, 2, \dots, r$ ) are known.

Taking into account eqn (16), after substitution of the values of  $\mu_{iK}$  and  $v_{iK}$  the algebraic system (17) for defining  $l_{1K}, l_{2K}, \dots, l_{rK}$  takes the form

$$\sum_{i=0}^{n_0} \left[ \left( \sum_{j=1}^r \mu_{iK} l_{jK} + v_{iK} \right) - a_{iK}^E \right] \mu_{iK} = 0 \quad (j = 1, 2, \dots, r) \quad (19)$$

#### Realization of Adjustable Parameters

##### Block-circuit with a Self-adjusting System using a Digital Computer

The duration of the intervals of constancy of the coefficients of reference eqn (3), when a digital computer is used in the control system, must satisfy correlation

$$t_{K+1} - t_K = T_1 + T_2 + T_3 + \Delta t \quad (20)$$

where  $T_1 = \Delta \tau (S - 1)$  is the time required to carry out measurements;  $T_2 = N/n_0$  is the time required for the computations;

$T_3$  is the time of actuator operation;  $0 \leq \Delta t \leq t_{K+1} - t_K$ ;  $\Delta \tau = \tau_{j+1} - \tau_j$  is the period of measurements ( $j = 1, 2, \dots, S$ );  $n_0$  is the computer speed of action, and  $N$  is the number of operations required to define coefficients  $l_{jK}$  ( $j = 1, 2, \dots, r$ ).

It is obvious that to ensure better operation of the self-adjusting system, it is necessary to reduce as much as possible the magnitude  $T = T_1 + T_2 + T_3$ .

Now the opportunities for reducing the time  $T_3$  are dealt with. This question is directly linked with the choice of the actuator. Electromechanical servosystems with a considerable time constant are usually employed as actuators at the present time. But it turns out that it is possible to suggest a number of purely circuit variants of the change of transfer functions or of gains of the correcting devices (regulators) of the system.

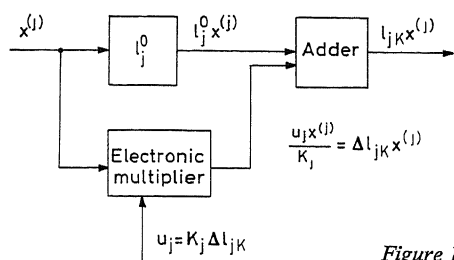


Figure 1

These inertia-less actuators are termed 'static'. It is particularly advantageous to produce static actuators with the aid of non-linear resistors (varistors), valves with variable gains (varimu), electronic multipliers, etc.

Consider, for example, one of the variants of a static actuator based on an electronic multiplier. Let the mode of control have the form

$$y = \sum_{j=1}^r l_j x^{(j)}$$

and let the  $j$ th adjustable parameter have the value  $l_j^0$  at moment  $t_0$  (start of operation of the system). While the system operates in accordance with the signals of the computer, the value  $l_j$  is constantly being corrected.

Thus, at the end of the interval  $(t_K, t_{K+1})$  one has

$$l_{jK} = l_j^0 + \Delta l_{jK} \\ y = \sum_{j=1}^r l_j^0 x^{(j)} + \sum_{j=1}^r \Delta l_{jK} x^{(j)} \quad (21)$$

Obviously each addend in the right-hand side of expression (21) can be instrumented with the aid of the circuit in Figure 1.

The following are self-adjusting system computer operating algorithms: when the real process is approximated by differential eqns (5), the algebraic systems (9), (18), and (19); when the real process is approximated by differential eqns (10), the algebraic systems (13), (18), and (19).

It is obvious that in the general case it is more convenient to solve the problem of self-adjustment according to the proposed principle with the aid of a high-speed digital computer. It can be specialized for solving systems of algebraic equations. Figure 2 shows the block diagram of a self-adjusting system with a digital computer.

### Some Particular Cases

In the preceding sections the proposed principle for creating a self-adjusting control system for non-stationary plants was expounded in general form. In practice, one may naturally encounter cases when the given principle can be used in more simplified variants. Several such opportunities are considered.

(1) Obviously, the entire theory expounded above can be applied fully to stationary and quasi-stationary systems, which are particular instances of non-stationary systems. In this case the durations of the intervals of constancy of the coefficients  $(t_K, t_{K+1})$  equal, for stationary systems

$$K=0, t_{K+1} - t_K = t_1 - t_0 = T_0 - t_0 \quad (22)$$

for quasi-stationary systems

$$t_{K+1} - t_K \geq \Delta t_p \quad (23)$$

where  $t_0$  is the control time (duration of the transient process).

As can be seen from relations (22) and (23), in stationary and quasi-stationary systems one is less rigidly confined to the time of analysis of the real process and synthesis of controller parameters. It is therefore possible to define coefficients  $a_{iK}$  and  $b_{iK}$  more accurately and to use criteria which reduce the self-adjustment process speed, but make it possible to increase the accuracy of operation of the system. Among such criteria one may cite, in particular, the integral criteria for the evaluation of the quality of a transient process<sup>3</sup>.

For stationary and quasi-stationary systems the problem of self-adjustment in accordance with the principle proposed above may be solved as a problem of the change in position of the roots of the transfer function of a closed system, i.e., the self-adjustment problem may be solved in accordance with the requirements of the root-locus method, which is extensively

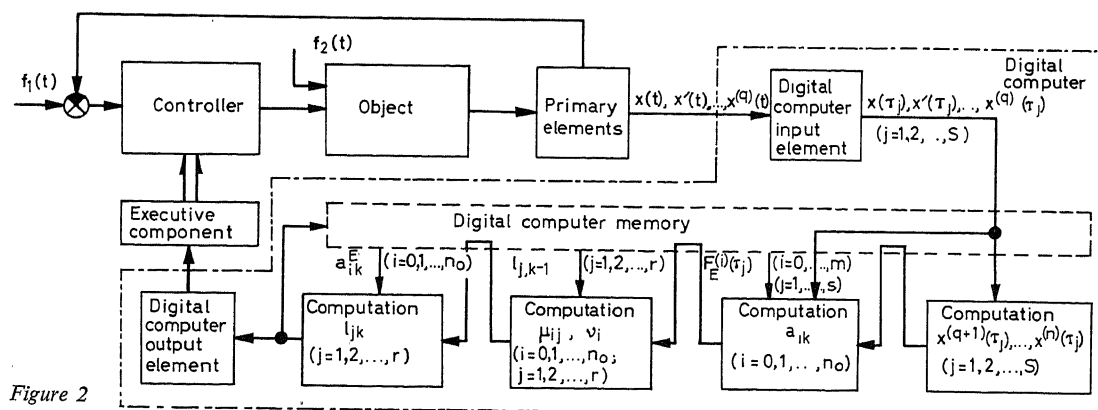


Figure 2

employed in automatic control theory. A feature of the use of the proposition of the root-locus method in accordance with the principle under consideration is that the zeros and poles defined by the coefficients  $a_{iK}$  and  $b_{iK}$  are fictions since they not only depend on the parameters of the controlled plant and controller, but also depend on real disturbances as well.

(2) In practice, one may encounter cases when a controller is required to ensure only the stability of a system in the course of operation. As is known, the stability of linear stationary systems is determined by the coefficients of the characteristic equation. This proposition is also valid for certain quasi-stationary systems (method of frozen coefficients).

Therefore to solve the problem posed (the provision of stability), the control system must define the actual values of the coefficients of the left-hand side of the differential equation of the system and must set on the controller such gains as will satisfy the conditions of stability, for example the conditions of the Hurwitz algebraic criterion. On the assumption that disturbance  $f$  is constant in the interval  $(t_K, t_{K+1})$  the coefficients of the characteristic equation of the system on this interval are determined in the following way.

The differential equation of the system for  $t \in (t_K, t_{K+1})$  is written in the form

$$x^{(n)} + \sum_{i=0}^{n-1} a_{iK} x^{(i)} = F_K$$

where  $F_K$  is in the general case the unknown right-hand side, constant for  $t \in (t_K, t_{K+1})$ . The algebraic system for determining the described coefficients will then be written thus:

$$x^{(n)}(\tau_j) + \sum_{i=0}^{n-1} x^{(i)}(\tau_j) a_{iK} = F_K \quad (j=1, 2, \dots, S) \quad (24)$$

Since  $F_K$  is unknown, but is constant in the interval  $(t_K, t_{K+1})$  it is eliminated with the assistance of one of the equations of system (24). For this purpose one uses the equation

$$x^{(n)}(\tau_l) + \sum_{i=0}^{n-1} x^{(i)}(\tau_l) a_{iK} = F_K \quad (1 \leq l \leq S)$$

After eliminating  $F_K$  one has:

$$\sum_{i=0}^{n-1} [x^{(i)}(\tau_j) - x^{(i)}(\tau_l)] a_{iK} = -[x^{(n)}(\tau_j) - x^{(n)}(\tau_l)] \quad (j=1, 2, \dots, l-1, l+1, \dots, S) \quad (25)$$

By resolving system (25) directly with  $S = n + 1$ , or by the least-squares method with  $S > n + 1$ , one determines the coefficients  $a_{iK}$  ( $i = 0, 1, \dots, n - 1$ ), the values of which are used, if the need arises, for synthesis of the values of the controller parameters which ensure the stability of the system.

## Conclusion

The paper has expounded only the basis of the proposed principle for the construction of a self-adjusting control system in general form and in certain particular cases. Studies are under way on problems connected with the approximation of differential equations with essentially variable coefficients by differential equations with piecewise-constant coefficients, with the selection of the type of computer to operate in the self-adjustment loop, with the dynamic precision of the self-adjusting system, etc. The investigations which have been made allow one to hope that the use of the principle expounded in this paper for the construction of self-adjusting control systems will prove extremely effective in many cases when it is expedient to use the natural oscillations of the system, without introducing test disturbance signals.

## References

- 1 GONCHAROV, V. L. *The Theory of Interpolation and Approximation of Functions*. 1954. Moscow; Gostekhizdat
- 2 LANCZOS, K. *Practical Methods of Applied Analysis*. 1961. Moscow; Fizmatgiz
- 3 POPOV, E. P. *Dynamics of Automatic Control Systems*. 1954. Moscow; Gostekhizdat
- 4 CHERNETSKII, V. I., and YUSUPOV, R. M. On one type of adaptive control system. *Izv. Akad. Nauk SSSR, Otdel Tekhn. Nauk* (1962)
- 5 VITENBERG, I. M. Specialized electric analog with automated scan. *Use of Computer Techniques for Production Automation*. 1961. Moscow; Mashgiz
- 6 POPOV, E. P. *Automatic Regulation and Control*. 1962. Moscow; Fizmatgiz

## DISCUSSION

K. S. P. KUMAR, *Control & Information Systems Lab., School of Electrical Engineering, Purdue University, Lafayette, Indiana, U.S.A.*

The authors have replaced a non-stationary system with an equivalent system with piece-wise constant coefficients. This is useful if the interval over which this approximation is valid is longer than the computation time needed to identify the coefficients. It would be interesting to hear of any experimental results on this.

Under the assumption of constant or piece-wise constant coefficients, an alternative procedure that is computationally very efficient and conceptually simple is presented below. The method consists in viewing the identification as a multi-point boundary value problem. A scalar example is given for illustrative purposes. Generalization to the vector-matrix case is very straightforward. The method does not need any test signals and derivatives of the input. Furthermore, the procedure is unchanged even if the system is non-linear.

Let the control system be described by

$$\dot{x} = fx + gu \quad (1)$$

where  $u$  is the control variable and  $f, g$  are unknown constants. As  $f$  and  $g$  are constants, describe them by

$$\dot{f} = 0 \quad (2)$$

$$\dot{g} = 0 \quad (3)$$

Make measurements on  $u$  and  $x$  over a finite time interval, the measurements on  $x$  serving as boundary values to solve the system of eqns (1) to (3). The method used to solve this boundary value problem is the quasi-linearization method<sup>1</sup>. The method has several advantages, chief among which is quadratic convergence. If, in addition to the coefficients, the order of the differential equation of the system is also unknown, the above method can still be successfully used to identify and control the system. The method can also be used for combined identification and control. These and other problems have been studied at Purdue University with great success and they will be reported very soon.

## Reference

- 1 KALABA, R. On non-linear differential equations, the maximum operation, and monotone convergence. *J. Math. Mech.* 8 (1959) 519

E. P. POPOV, *in reply*

The method suggested by Dr. Kumar is close to that described in our paper. We are now working on problems of identification of non-linear systems.

S. S. L. CHANG, *College of Engineering, State University of New York, Long Island, N.Y., U.S.A.*

I have two particular questions to ask:

(1) In the paper the set of coefficients  $a_{ik}$  and  $b_{ik}$  are assumed to be independent for each sub-interval  $K$ . Sometimes the unknown disturbances are quite considerable and the coefficients  $a_{ik}$  and  $b_{ik}$  cannot be accurately determined for each sub-interval as the authors proposed. An alternative is to use the newly measured data to modify previous values of  $a_{ik}$  and  $b_{ik}$ . Have the authors studied this possibility and, if so, have they any comments to make?

(2) Will the authors please comment on the relative merits of their proposed method versus the dual control method of Professor Feldbaum?

E. P. POPOV, *in reply*

Professor Chang suggests, as we understand it, that coefficients  $a_{ik}$  and  $b_{ik}$  should not be determined directly on each sub-interval ( $t_K, t_{K-1}$ ), but that their previous values  $a_{i, K-1}$  and  $b_{i, K-1}$  should be clarified by processing the new information on the interval ( $t_K, t_{K-1}$ ), i.e., the quantities  $\Delta a_{ik} \Delta b_{ik}$  should be determined, in order to use them subsequently for determining the coefficients

$$a_{iK} = a_{i, K-1} + \Delta a_{iK}$$

$$b_{iK} = b_{i, K-1} + \Delta b_{iK}$$

We have considered such algorithms. It proved difficult to use them because of the impermissibly high values of the absolute errors of quantities  $\Delta a_{ik}$  and  $\Delta b_{ik}$  due to the approximation, and also the existence of measuring and computing errors when the values of the quantities  $\Delta a_{iK}$  and  $\Delta b_{iK}$  themselves are quite small.

As far as comparison of our adaptive system with dual control systems is concerned, it must be said that in our opinion any adaptive system, including ours, is a dual-control system.

F. B. TUTEUR, *Yale University Department of Engineering and Applied Science, New Haven 11, Connecticut, U.S.A.*

(1) The authors propose identification of parameters by the use of high-order derivatives of the measured output quantities. However, it is well known in practice that derivatives of almost any order are almost impossible to obtain with any accuracy from measured data, because the process of differentiation simply amplifies the noise. Thus

this method does not seem very practical. Use of a digital computer is of no help in this respect either. On the other hand, the proposal to substitute high-order integration would probably suffer from the disadvantage that important variations of the signal would be smoothed out by the integration process. The authors' comments on this point would be appreciated.

(2) Since the authors propose to treat a non-stationary system by breaking it up into a sequence of piece-wise stationary systems, the important question arises regarding how large the true 'interval'  $t_k, t_{k+1}$  should be made. If it is too short eqn (20) of the paper is not satisfied, and if it is too long the system cannot be considered to be stationary. In this connection it should be pointed out that the time required to identify a parameter is, in general, several orders of magnitude greater than the significant system time constants<sup>1, 2, 3</sup>. It would seem, therefore, that the method could only be used for very slowly-varying systems.

#### References

- <sup>1</sup> BLANDHOL, E. and BALCHEN, J. G. Determination of system dynamics by means of adjustable models. *2nd I.F.A.C. Congr.* Basle. 1963. London; Butterworths: Munich; Oldenbourg
- <sup>2</sup> QVARNSTRÖM, B. An uncertainty relation for linear mathematical models. *2nd I.F.A.C. Congr.* Basle. 1963. London; Butterworths: Munich; Oldenbourg
- <sup>3</sup> MESCH, F. A comparison of measuring time in self-adjusting control systems. *2nd I.F.A.C. Congr.* Basle 1963. London; Butterworths: Munich; Oldenbourg

E. P. POPOV, *in reply*

We are in full agreement with Professor Tuteur's comments on the difficulties which arise because of the need to have derivatives of a controlled variable of high order. We have studied, and are studying, this problem.

Some algorithms of the solution of the problem of identification with only a limited number of derivatives are presented in the paper. They had to be dealt with very briefly because of lack of time.

Greatest attention should be paid to the algorithms obtained on the basis of multiple integration of differential equations, each of which characterizes one degree of freedom, i.e. has an order not higher than second.

We can reply to Professor Tuteur's second comment as follows.

For specific systems, selection of the length of sub-interval  $[t_k, t_{k+1}]$  was performed on an analogue computer, proceeding from the requirement of the specified accuracy of approximation of the solution of the differential equation with variable coefficients describing the system by solving an equation with constant coefficients.

The coefficients  $a_{ik}$  and  $b_{ik}$  were determined on the digital and on the analogue computer from the results of measurements at intervals, the length of which was less than the essential time constants of the system.

# On the Dynamics of Some Learning and Self-Learning Processes

V.K. CHICHINADZE

## Summary

The paper deals with the question of the use of self-learning processes in adaptive systems with least initial information. The mathematical expression, based on the phenomenon of entropy by which the dynamics of the learning and self-learning processes are experimentally determined, is given in the paper. Also introduced is the phenomenon of rate of learning which can give the value of the processes. Some concrete examples of the learning of automata and some living beings are discussed. In the paper are given some experimental data of the results of the self-learning system making the automatic synthesis of the corrective device. According to the experimental data the processes of searching can be significantly lessened by using the methods of learning of the adaptive system. The proposed mathematical expression can be used for the description of different learning processes and makes it possible to achieve some analogies between them. Experiments connected with the self-learning of the adaptive system were carried out on the special electronic machine of the Institute of Electronics, Automatics and Telemechanics of the Academy of Sciences of Georgia.

## Sommaire

Cette note concerne l'utilisation des processus à auto-apprentissage dans les systèmes adaptatifs disposant d'un minimum d'information initiale. Elle donne, en se fondant sur l'entropie, l'expression mathématique à l'aide de laquelle la dynamique des processus d'apprentissage et d'auto-apprentissage peut être étudiée. Elle introduit la notion de taux d'apprentissage, qui peut caractériser la valeur de ces processus. Elle donne quelques exemples concrets d'apprentissage dans les automates et chez quelques êtres vivants, ainsi que des données expérimentales relatives aux résultats obtenus sur des systèmes à auto-apprentissage réalisant la synthèse automatique de systèmes correcteurs. D'après ces données expérimentales, il apparaît que le processus de recherche peut être nettement réduit en utilisant la méthode d'apprentissage des systèmes adaptatifs. L'expression mathématique proposée permet de décrire différents processus d'apprentissage et d'établir des analogies entre eux. Les expériences relatives à l'auto-apprentissage des systèmes adaptatifs ont été conduites à l'aide de la machine électronique spéciale de l'Institut d'Electronique, d'Automatique et de Télémécanique de l'Académie des Sciences de Georgie.

## Zusammenfassung

Dieser Beitrag befaßt sich mit der Frage der Verwendung selbstlernender Prozesse in anpassenden Systemen mit minimaler Anfangsinformation. Der auf dem Phänomen der Entropie beruhende mathematische Ausdruck, anhand dessen die Dynamik des Lern- und Selbstlernprozesses gesucht wird, wird angegeben. Auch wird die Lerngeschwindigkeit angeführt, die einen Wertmaßstab für die Prozesse abgeben kann. Einige konkrete Beispiele für das Lernen in Automaten und Lebewesen werden besprochen. In diesem Beitrag werden einige Angaben über die Versuchsergebnisse eines Selbstlernsystems gegeben, das die automatische Synthese eines Korrekturgliedes durchführt. Mit Hilfe der Lernmethoden für das anpassende System können, wie diese Daten zeigen, die Suchvorgänge wesentlich verkürzt werden. Der hier vorgeschlagene mathematische Ausdruck kann für die Beschreibung verschiedener Lernverfahren benutzt werden und macht Vergleiche zwischen ihnen in gewisser Beziehung möglich. Die Ver-

suche im Zusammenhang mit dem Selbstlernen des anpassenden Systems wurden auf einen besonderen Elektronenrechner des Institutes für Elektronik, Automatik und Telemechanik der Georgischen Akademie der Wissenschaften durchgeführt.

## Introduction

At the present time, side by side with the existing usual control systems, the so-called adaptive systems, which are themselves capable of defining the necessary structure in a given situation, are being worked out. As a rule the definition of a new structure in the adaptive system is performed with the help of the algorithm put in beforehand.

Of great interest is the following step in the development of the adaptive system, when a device without a given algorithm will be capable of performing all the necessary changes according to the stored information.

These systems will be different from the usual ones, working on the definite algorithm of search by the degree of determination. The solution of this problem can be found by the application of self-learning methods. The adaptive system with self-learning properties will be generally characterized by the capacity of the initial information which has been put into the system by the designer. Devices working with the minimum initial information are, for example, the perceptron created by Rosenblatt and the homeostat first proposed by Ashby. If the perceptron has self-learning properties and by storage of definite knowledge can decrease the initial uncertainty, Ashby's homeostat lacks such properties. The self-learning process can be fulfilled after a special storage device is added to the homeostat, when the system acquires the properties of self-organization.

## Dynamics of Learning Processes

While examining the systems with the minimum initial information and the subsequent increase of information by self-learning processes, there arises the question of the measure of the initial uncertainty at which the system starts its organization.

This uncertainty must decrease while the number of tests are increased. The measure of such uncertainty must be based on the phenomenon of entropy. But the latter needs some changes. Actually the entropy has its maximum value when all the events are equally probable. When the processes of self-learning or learning take place, the entropy of the system at the beginning of the process may have minimum or even zero value, and although according to Shannon the system will be fully determined, it may be on its lowest step of organization. This is because the probability of the event which the system learns has a smaller value than the probabilities of other events.

As we are interested in the process of organizations of the system, followed by the decrease of the vagueness, the term is

introduced to differentiate it from Shannon's uncertainty. Let us take some function  $V$  as the measure of the latter.

$$V = \begin{cases} 2H_M - H & S < S_M \\ H & S \geq S_M \end{cases} \quad (1)$$

where  $H_M$  is the maximum value of the entropy,  $H$  is the entropy of the system,  $S$  is the number of trials, and  $S_M$  is the trial when the entropy acquires its maximum value.

As is known

$$H = - \sum_{i=1}^n P_i \log P_i.$$

The entropy acquires its maximum value when all the  $n$  events are equally probable

$$P_1 = P_2 = \dots = P_n = \frac{1}{n}$$

then

$$H_M = -n \frac{1}{n} \log \frac{1}{n} = \log n \quad (2)$$

The last expression corresponds to Hartley's information capacity. Let us define it as  $C_1$ . Hence, the function of vagueness

$$V = \begin{cases} 2C_1(n) - H & S < S_M \\ H & S \geq S_M \end{cases} \quad (3)$$

So the initial vagueness of the system will be equal to

$$V_0 = C(n) - H_0 \quad (4)$$

where  $C(n)$  is the doubled Hartley's information capacity and  $H_0$  is the initial entropy of the system which is determined according to the probability of different events at the beginning of the learning process.

The maximum initial vagueness of the system is equal to

$$V_{0M} = C(n) \quad (5)$$

and will take place when the initial entropy of the system  $H_0 = 0$ . In spite of the fact that the system according to Shannon is fully determined, the initial vagueness of the system takes its maximum value.

Let us introduce some non-linear function  $\lambda(S)$

$$\lambda(S) = \begin{cases} 1 & S < S_M \\ \frac{C(n)}{H} - 1 & S \geq S_M \end{cases} \quad (6)$$

Eqn (3) will have the final form

$$V = C(n) - \lambda(3) H(P_1, P_2 \dots P_n) \quad (7)$$

According to the above, the evaluation of the degree of vagueness is carried out by means of the function

$$V = f(P_1, P_2 \dots P_n) \left( \sum_{i=1}^n P_i = 1 \right) \quad (8)$$

where  $P_i$  is the probability of the appearance of an event which gives the acceptable solution of the problem.

The learning is performed by a number of cycles of trials in the presence of some stimuli in the case of living beings or the meaning of the environment in the case of automatic control.

After each cycle a new value of  $V$  is received, according to the redistribution of  $P$  probabilities. Hence, there is an expression

$$V = \varphi(S) \quad (9)$$

where  $\varphi$  determines the learning process.

Let us use the difference equation for the description of the learning process.

Consider that  $V(S)$  is determined for the number of values  $S + C\Delta S$ , where  $C$  has the value 0, 1, 2, ...,  $m$ . As is known, the difference of the first order  $\Delta V$  will be defined as

$$\Delta V_S = V_{S+\Delta S} - V \quad (10)$$

the difference of the second order:

$$\Delta^2 V_S = \Delta V_{S+\Delta S} - \Delta V_S \quad (11)$$

Analogically the difference of  $K$ th order:

$$\Delta^K V_S = \Delta^{K-1} V_{S+\Delta S} - \Delta^{K-1} V_S \quad (12)$$

On the other hand

$$\Delta^K V_S = \sum_{\tau=0}^K (-1)^\tau \frac{K}{\tau!(K-\tau)!} V_{S+n-\tau} \quad (13)$$

Consider that the learning process, accompanied by the change of  $V$ , when  $S$  is increasing, is described by a non-linear difference equation, which is a non-linear relation of unknown  $V$  and its differences. The equation will have the form:

$$\beta_n \Delta^n V_S + \beta_{n-1} \Delta^{n-1} V_S + \dots + \beta_1 \Delta V_S + \beta_0 V_S = 0 \quad (14)$$

where  $\beta$  depends on  $V$  and its differences.

This equation may have a stable, non-stable or oscillatory solution and according to this the system may be considered as a learned one or the learning process may be unsuccessful. When the stable process takes place the following condition must be fulfilled:

$$\lim_{S \rightarrow \infty} V = 0 \quad (15)$$

The difference equation may be substituted by some non-linear equation

$$V^\tau = f(V, V', V'' \dots V^{(\tau-1)}) \quad (16)$$

which may also have different solutions.

The main characteristic feature of the learning process may be considered being the decrease of  $V$  vagueness which is characterized by  $V$  function. The reason for the latter can be found in the decrease of the probability of the event which is the most essential for the given learning process. Such event may be the recognition of a definite learning pattern, the determination of a control system, etc.

The velocity of the learning process which conditionally will be called the rate  $\delta$  is the significant factor which presents the difference of the first order

$$\delta = V_{(S+\Delta S)} - V_{(S)} \quad (17)$$

For a continuous process

$$\delta = \frac{dV}{dS} \quad (18)$$

According to eqn (3)

$$\delta = \begin{cases} -\frac{dH}{dS} & S < S_M \\ \frac{dH}{dS} & S \geq S_M \end{cases} \quad (19)$$

The increase of  $V$  function is determined by

$$\Delta V = \begin{cases} \sum_{i=1}^n P_{i(S)} \log P_{i(S)} - \sum_{i=1}^n P_{i(S-1)} \log P_{i(S-1)} & S < S_M \\ -\sum_{i=1}^n P_{i(S)} \log P_{i(S)} + \sum_{i=1}^n P_{i(S-1)} \log P_{i(S-1)} & S \geq S_M \end{cases} \quad (20)$$

where  $P_{i(S)}$  is the value of probability of  $i$  events in  $S$  trials and  $P_{i(S-1)}$  is the value of probability of  $i$  events in  $(S-1)$  trials. Hence, the rate of learning

$$\delta = \begin{cases} \lim_{\Delta S \rightarrow \infty} \frac{\sum_{i=1}^n P_{i(S)} \log P_{i(S)} - \sum_{i=1}^n P_{i(S-1)} \log P_{i(S-1)}}{\Delta S} & S < S_M \\ \lim_{\Delta S \rightarrow \infty} \frac{\sum_{i=1}^n P_{i(S-1)} \log P_{i(S-1)} - \sum_{i=1}^n P_{i(S)} \log P_{i(S)}}{\Delta S} & S \geq S_M \end{cases} \quad (21)$$

The rate of learning may be a criterion of the quantitative evaluation of the learning process. From this point of view it is convenient to analyse the learning process on the phase plane, the coordinates of which are the rate of learning  $\delta$  and the value of vagueness. In Figure 1(b) are given several trajectories corresponding to the stable learning process (curve 4), to the non-stable process (curve 2) and to the oscillatory process (curve 3).

In some cases the learning process can be finished at small values of the vagueness function, i.e. when

$$\lim_{S \rightarrow \infty} V = \varepsilon \quad (22)$$

where

$$\varepsilon \ll C \quad (23)$$

According to eqn (7) the learning process, accompanied by the decrease of  $V$  vagueness, takes place because of the increase of one of  $P_{(i)}$  probabilities which corresponds to the learning event and the decrease of all other probabilities. The functional relations

$$P_i = f(S) \quad (24)$$

are different, and as experiments show are mainly determined by exponential laws.

Later, the learning process of the adaptive system, with the minimum initial information at which the probability changes according to the exponential function, will be analysed.

### Concrete Examples of Learning Processes

Let us examine several examples of learning processes with the aim of drawing analogies between them. As already mentioned, the perceptron is a device with the minimum initial information. It is assumed that at the beginning of work prob-

abilities of recognition of each recognition pattern are equal. The  $V$  function in this case is equal to

$$V(0) = -P_\psi \log P_\psi - \sum_{i=1}^{n-1} P_i \log P_i \quad (25)$$

where  $P_\psi$  is the probability of recognition of any one pattern the perceptron will learn.

The above assumption, meaning

$$P_1 = P_2 = \dots = P_\psi = \dots = P_n \quad (26)$$

cannot be characteristic for all the processes. Sometimes the probability of appearance of an event connected with learning is smaller than other probabilities.

Each trial in the perceptron, connected with placing on the retina the same pattern, is followed by  $P_\psi$  increase and  $P_i$  ( $i = 1, 2, \dots, n-1$ ) decrease. The changing law of probabilities may be considered as exponential

$$\begin{aligned} P_\psi &= 1 - \frac{n-1}{n} e^{-\alpha S} \\ P_i &= \frac{1}{n} e^{-\alpha S} \end{aligned} \quad (27)$$

where  $\alpha$  is a constant. After the first trial  $V(S)$  according to eqn (7) will have the form

$$\begin{aligned} V_1(S) &= -\left(1 - \frac{n-1}{n} e^{-\alpha}\right) \log \left(1 - \frac{n-1}{n}\right) \\ &= e^{-\alpha} - \frac{n-1}{n} e^{-\alpha} \log \frac{e^{-\alpha}}{n} \end{aligned} \quad (28)$$

After the second trial:

$$\begin{aligned} V_2(S) &= -\left(1 - \frac{n-1}{n} e^{-2\alpha}\right) \log \left(1 - \frac{n-1}{n}\right) \\ &= e^{-2\alpha} - \frac{n-1}{n} e^{-2\alpha} \log \frac{e^{-2\alpha}}{n} \end{aligned} \quad (29)$$

Analogically after  $S$  trials

$$\begin{aligned} V_S(S) &= -\left(1 - \frac{n-1}{n} e^{-S\alpha}\right) \log \left(1 - \frac{n-1}{n}\right) \\ &= e^{-S\alpha} - \frac{n-1}{n} e^{-S\alpha} \log \frac{e^{-S\alpha}}{n} \end{aligned} \quad (30)$$

The rate of learning for the perceptron is defined by

$$\begin{aligned} \delta &= \frac{n-1}{n} e^{-\alpha S} \log \left( \frac{e^{-\alpha S}}{n} + 1 \right) \\ &\quad - \frac{\alpha(n-1)}{n} e^{-\alpha S} \log \left( 1 - \frac{n-1}{n} e^{-\alpha S} \right) \end{aligned} \quad (31)$$

In Figure 1(a) is given the learning process of pattern recognition. In Figure 1(b) (curve 1) is given the phase trajectory of the above-mentioned process. The curves are drawn for  $n = 20$  and  $\alpha = 0.1$ .

As seen in Figure 1(a) for 50 cycles of trials the vagueness of the system in the learning process decreases from  $V_{(0)} = 4.32$

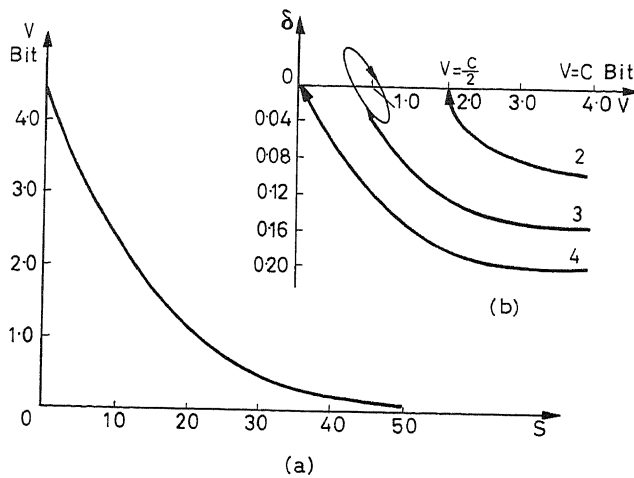


Figure 1

bit to  $V = 0.12$  bit, i.e. by 36 times. By further increase of the number of cycles  $V$  can be decreased to the arbitrary small value.

Now consider the learning process for reproduction of words. The process is based on the data obtained in the experiments made by Bruner and Zimmerman<sup>1</sup>. The experiment consists in the following: a list of 32 mono-syllable words is read to the person being taught. The person then writes down those words which he has remembered. Then, to avoid correlation, the order of words is changed and the process is repeated.

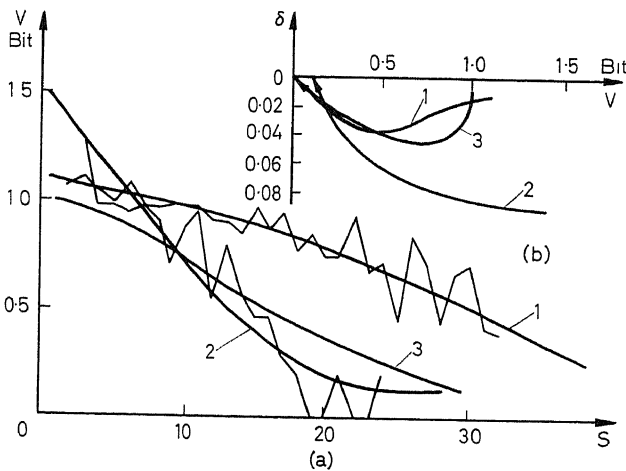


Figure 2

In Figure 2(a) is given a curve of learning process (curve 1) according to eqn (7).  $P(S)$  is determined according to the experimental data. In this case there are two events: (1) the remembering of the whole table, and (2) the case of non-remembering.  $V$  is determined by the probabilities  $P$  and  $1 - P$ . In Figure 2(b) curve 1 corresponds to the phase trajectory of the process. As seen in Figure 2(a) the experiment was finished, when  $V_K(S) = 0.4$  bit. The initial vagueness is equal to  $V_M(0) = 1.1$  bit. In this case the learning process cannot be considered as fulfilled, because the relation  $V_M(S)/V_K(S)$  is sufficiently small. But the extrapolation of the curve allows the conclusion that when  $S = 50$ ,  $V_K(S) = 0.02$  bit. In this case the vagueness decreases by 55 times.

Of interest are the experiments effected by Solomon and Win. In one experiment a dog is taught to avoid pain which is the result of a current. According to the data given, the corresponding learning process calculated by eqn (7) is shown in Figure 2(a) (curve 2).

In Figure 2(b) is given the corresponding phase trajectory (curve 2) for which the equilibrium point is not at the origin of the coordinates. This means that after the learning process is finished, vagueness equal to

$$\lim_{S \rightarrow \infty} V(S) = \varepsilon \quad (32)$$

is left.

The learning process of a hungry rat which is placed into a T-shaped labyrinth is similar. In Figure 2(a) and 2(b) curve 3 shows the character of changes of  $V$  vagueness with increase of the number of cycles of trials and the phase trajectory of the process. In this case the vagueness decreases by 12.5 times in 30 trials. The phase trajectory is compressed to the origins of the coordinates and

$$\lim_{S \rightarrow \infty} V(S) = 0 \quad (33)$$

#### The Learning Process of Adaptive System Fulfilled by the Exclusive Method

Consider the self-learning process of the adaptive system working on minimum initial information. As is known, such a system defines some subset  $M$  by random search. The elements of the subset are the points corresponding, from some point of view, to the position of any vector  $Y$ , which characterizes the state of the system. If  $N$  is the set of all possible states, then  $M \subset N$ . According to the above, when  $Y$  is the domain defined by  $M$  subset, the search process and the learning process of the adaptive system must be finished. First, the direction of  $Y$  is defined according to the law of random numbers, because of the lack of some information. Then the evaluation of the given domain takes place and if the latter does not correspond to requirements, the coordinates of the vector domain will be stored by a special memory device. In the following cycles of trials  $Y$  cannot return to the already tested position corresponding to the set. Thus the exclusion of one element from  $N$  set takes place. Then the system acts analogically. Assume that  $N$  and  $M$  are restricted final sets. By  $m$  and  $n$  designate the power of final sets of  $M$  and  $N$  correspondingly. As  $P$  probability of any  $Y$  position is equal to  $P = 1/n$ ,  $P_p$  probability of  $Y$  position defined by  $M$  set will be equal to

$$P_p = \sum_{i=n-m}^n P_i = \frac{m}{n} \quad (34)$$

The value of probability in the following cycles will be

$$P = \frac{1}{n-S} \quad P_m = \frac{m}{n-S} \quad (35)$$

In the considered adaptive system the direction of the vector in the presence of storage is defined by the amplifier gain in four different channels each of which can obtain seven discrete values. The total number of sets of the system is equal to  $n = 7^4 = 2,401$ . In the adaptive system the process of self-learning begins at the smaller values of information capacity,



as  $m > 1$  and  $P_m > P_i$ . The information capacity coinciding with the value of the initial vagueness of the system is defined as

$$V_{(0)} = -P_4 \log P_4 - \sum_{j=1}^{n-m} P_j \log P_j \quad (36)$$

The vagueness of the adaptive system after  $S$  cycles will be equal to

$$V_{(S)} = -\frac{m}{n-S} \log \frac{m}{n-S} - \sum_{j=1}^{n-S} \frac{1}{n-S} \log \frac{1}{n-S} \quad (37)$$

After some transformations

$$V_{(S)} = -\frac{m}{n-S} \log m + \log(n-S) \quad (38)$$

The rate of learning

$$\delta = \frac{dV}{dS} = -\frac{m}{(n-S)^2} \log m - \frac{1}{n-S} \quad (39)$$

Eqn (39) shows that the rate of learning increases in absolute value with the increase of number of  $S$  trials. Since  $0 < S \leq n-m$ , the rate of learning reaches its maximum at the end of the learning process when

$$S = n - m$$

According to eqns (38) and (39),  $V$ , the vagueness of the system, and  $\delta$ , the rate of learning, depend also on the power of sets  $M$  and  $N$ .

Let us consider now some concrete examples of work of the adaptive system when the adaptive systems automatically define the structure and the value of parameters of correcting devices. The domain of the adaptive system is evaluated by the well-known integral criterion.

The learning process consists in the following: the model of an object, the equation of which is known, is connected to the adaptive system, which changes its structure and parameters till satisfactory work of the whole system is accepted. The storage device does not allow the repetition of the former state. Such a process, which can conventionally be called the learning process, by the exclusion method does not essentially differ from the analysed learning process of a dog and a rat.

The results of the experiments of the work of the adaptive systems by the definition of the required structure of the corrective devices are given below. Here is an automatically controlled plant which is described by the equations:

$$\begin{aligned} x_1 &= 30z - x_3 - x_4 \\ 0.025x_2'' + 0.125x_2' + x_2 &= 4x_1 \\ x_3' + x_3 &= 1.5x_2 \end{aligned}$$

The equation of the corrective device in the presence of which

$$\int [\alpha_1 x_3^2 + \alpha_2 (x_3')^2] dt \leq A$$

must be determined.

The adaptive system without the use of a memory device could not define in 8 h the corresponding structure. The search was fulfilled using self-learning methods and after 1,503 cycles of trials, taking 5 h 12 min, the structure of corrective device described by the equation:

$$0.01845x_4' + x_4 = 0.0063x_2' + 0.282x_2$$

was defined.

Repeating the search process, and having experience in the learning process, the adaptive system defined the new domain of the system, satisfying the given conditions, in 14 min after 23 cycles of trials. In Figure 3 (curve 1) is given the learning process computed according to eqn (37). In Figure 4 (curve 1) is given the phase trajectory of the learning process, defined according to eqn (39). In the given example only 12 domains from 240 satisfy the given conditions, i.e.  $m = 12$ . For the second example let us take the process of definition of the structure of parameters of a corrective device intended for some plant, described by

$$0.0005x_1''' + 0.051x_1'' + 1.06x_1' + 23x_1 = 22(z - x_2)$$

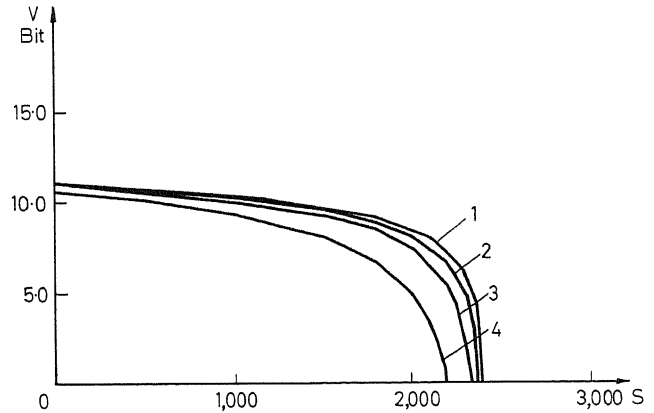


Figure 3

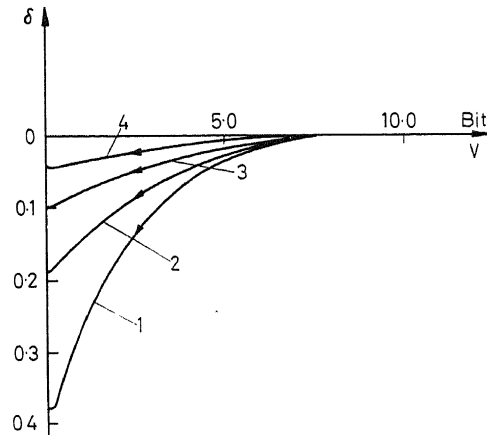


Figure 4

In Figure 3 (curve 2) is given the corresponding self-learning process. In Figure 4, curve 2 shows the corresponding phase trajectory. The equation of the corrective device, found as a result of self-learning, has the form

$$0.0001x_2'' + 0.01x_2' + x_2 = 0.001x_1'$$

As in this case  $m = 32$ , the self-learning process is fulfilled a little quicker. Curve 3 in Figure 3 corresponds to the self-learning process of the system, when the initial vagueness is defined by  $m = 72$ . The initial equation of the plant is

$$28.06x_1' + x_1 = 8.98(z - x_2 - x_3) \quad x_2' = x$$

The equation of the corrective device found by the adaptive system as a result of the learning process has the form

$$0.666 x_3'' + x_3' + x_3 = 0.666 x_1' + 0.666 x_1$$

The phase trajectory of the process is given in Figure 4 (curve 3).

Let us analyse the last example, which corresponds to the relatively small value of the initial vagueness, when  $m = 194$ . The curve of the learning process is given in Figure 3 (curve 4). The phase trajectory is given in Figure 4 (curve 4). The equations of the plant have the form

$$0.15 x' + x_1 = 2(z - x_4)$$

$$0.1 x_2' + x_2 = 2x_1$$

$$x_3' + x_3 = 50(x_2 - x_5)$$

$$1.5 x_4' + x_4 = x_3$$

The equation found by the adaptive system has the form

$$x_5'' + 1.3 x_5' + 0.36 x_5 = 0.1 x_3' + 0.3 x_3$$

The phase trajectory of the learning process of the adaptive system given in Figure 4 and also eqns (38) and (39) shows that the  $\delta = \gamma(V)$  function in  $V = 0$  point has a discontinuity of the first order when in the analysed examples the rate of learning was smoothly decreasing. The reason is the fact that learning by the exclusion method with the accepted assumptions is accompanied by no essential decrease of the vagueness at the beginning of the learning process and increase of velocity with the storage of information. The above follows the relation  $P = m/n - S$ . The opposite applies using other learning methods. At the beginning of the analysed process the decrease of rate of vagueness is relatively higher than the exponential velocity decrease. The self-learning process is based on the assumption that  $Y$  in the searching process does not enter the domain which is determined by  $M$  subset. Such a restriction is a strong one. If the process can be considered without such a restriction, then for the determination of the law of changing of the probabilities for  $Y$  entering this domain is determined by the hypergeometric law. The probability of  $Y$  passing some points  $m_K$  ( $0 \leq m_K \leq m$ ) will be determined by

$$W_K = \frac{\binom{m}{m_K} \binom{n-m}{S-m_K}}{\binom{n}{S}} \quad (40)$$

Hence for each  $S$  value the probability  $W_1$ , for  $Y$  entering the  $M$  domain can be determined and

$$W_K = \sum_{k=1}^m W_K \quad (41)$$

According to eqns (7), (40) and (41), giving different values for  $S$  can be determined by the function

$$W = f(S) \quad (42)$$

In Figure 5 (a) are given the curves of the self-learning process by the exclusive methods computed according to eqn (7) for the cases when  $n = 52$ ,  $m = 4$  (curve 1) and  $n = 10$ ,  $m = 2$  (curve 2).

In Figure 5 (b) are given the corresponding phase trajectories.

The data received as a result of experiments shows that the exclusive method can be used for cases when the initial vagueness has a sufficiently large value. The advantage of the exclusive method is that the work can be fulfilled without any search algorithm.

The experiments on self-learning were effected on a specialized computer which was designed at the Institute of Electronics, Automatics and Telemechanics of the Academy of Sciences of Georgia.

#### References

- BUSH, R. and MOSTELLER, F. *Stokhasticheskie Metodi Obucheniya*. Gosudarstvennoe Izdatelstvo Fiziko-Matematicheskoi Literaturi. Moscow. 1962
- ROSENBLATT, F. *Principles of Neurodynamics*. Spartan Books, 6411 Chillum Place, N.W. Washington, D.C.
- CHICHINADZE, V. K. O Nekotorikh Voprosakh Postroeniya Samonastroyayushchikhsya i Samoobuchayushchikhsya Sistem Avtomaticheskogo Upravleniya, Osnovauiikh na Printsipakh Sluchaiinogo Poiska. Trudui 1 Mezhdunarodnogo Kongressa Mezhdunarodnoii Federatsii po Avtomaticheskomu Upravleniyu. Vol. 2. Izdatelstvo Akademii Nauk SSSR. Moscow. 1961
- FELLER, V. *Vvedenie v Teoriyu Veroyatnostei i Ee Prilozhnie*. Izdatelstvo Inostrannoi Literaturi. Moscow. 1952

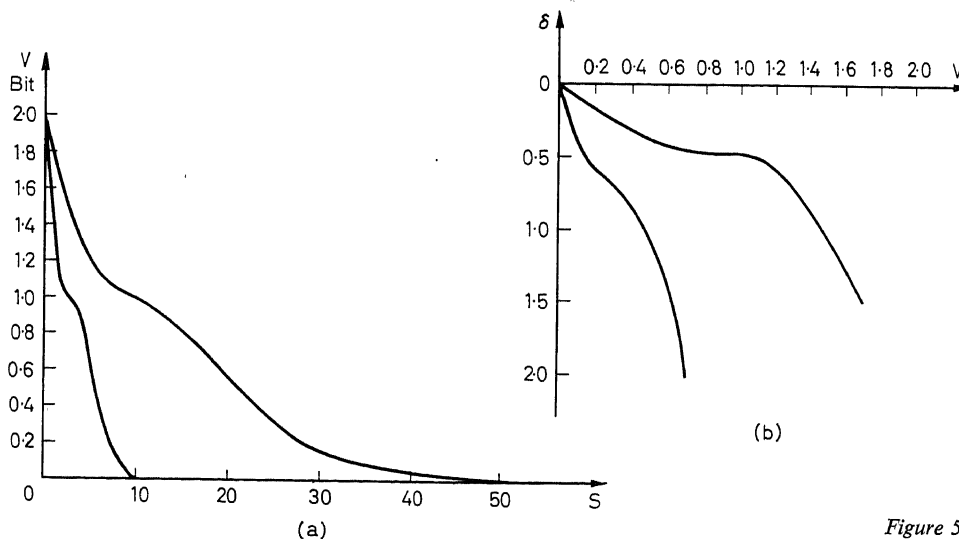


Figure 5

# On the Searching of Extrema of Functions in Automatic Control Systems

A. A. VORONOV and M. B. IGNATJEV

## Summary

A peculiar approach to the problem of searching extrema of functions of many variables on the intersection of multidimensional surfaces is considered. The differential equations the solutions of which are set on these multidimensional surfaces (manifolds), or on their intersections, are given. In the structure of these equations the arbitrary functional coefficients  $u_s$ , which determine the trajectory on the surfaces, are revealed. In case of reproducing the manifold  $F(x_1, x_2, \dots, x_n) = 0$ , the number of arbitrary coefficients is equal to  $C_n^2$ ; in case of the trajectory lying on the intersection of  $m$  manifolds  $F_j(x_1, x_2, \dots, x_n) = 0$ ,  $j = 1, 2, \dots, m$ , the number of arbitrary coefficients is equal to  $C_n^{m+1}$ .

Having certain concrete arbitrary coefficients  $u_s$  one can make devices performing the continuous searching of extrema of functions of many variables, the extremum point being the steady-state point for the structures considered. By choosing different coefficients it is possible to obtain the movement towards the extremum point along various trajectories (in particular along the geodetic curve). For example the differential equations, the trajectories of which are located on the surfaces  $F(x, y, z) = 0$  are:

$$\frac{dx}{dt} = u_1 \frac{\partial F}{\partial y} - u_2 \frac{\partial F}{\partial z}, \quad \frac{dy}{dt} = -u_1 \frac{\partial F}{\partial x} + u_3 \frac{\partial F}{\partial z}, \quad \frac{dz}{dt} = u_2 \frac{\partial F}{\partial x} - u_3 \frac{\partial F}{\partial y}$$

and the movement towards the maximum of  $z$  takes place when

$$u_2 = a_1^2 \frac{\partial F}{\partial x}, \quad u_3 = -a_2^2 \frac{\partial F}{\partial y}$$

The considered structures may be used for optimization of a composite parameter for complex systems, for solving the problem of construction of robot's motion, and for construction of ultra-stability systems.

## Sommaire

On examine une solution particulière pour résoudre le problème de recherche d'extrême de fonctions de variables multiples aux intersections des surfaces multidimensionnelles. On donne les équations différentielles dont les solutions se trouvent sur ces surfaces multidimensionnelles ou à leurs intersections. On localise les coefficients arbitraires fonctionnels  $u_s$  qui déterminent la trajectoire sur les surfaces. Dans le cas de la polyvalence  $F(x_1, x_2, \dots, x_n) = 0$ , le nombre de coefficients arbitraires est égal à  $C_n^2$ . Dans le cas où la trajectoire se trouve à l'intersection de  $m$  surfaces  $F_j(x_1, x_2, \dots, x_n) = 0$ ,  $j = 1, 2, \dots, m$ , ce nombre est égal à  $C_n^{m+1}$ .

Avec certaines valeurs concrètes des coefficients arbitraires  $u_s$ , on peut construire des dispositifs pouvant effectuer la recherche continue d'extrême de fonctions de variables multiples, le point extrême représentant l'état stationnaire des structures considérées. En choisissant convenablement ces coefficients, on peut réaliser le mouvement vers le point extrême selon des trajectoires variées (notamment le long de la géodésique). Par exemple, les équations différentielles dont les surfaces  $F(x, y, z) = 0$ , sont:

$$\frac{dx}{dt} = u_1 \frac{\partial F}{\partial y} - u_2 \frac{\partial F}{\partial z}, \quad \frac{dy}{dt} = -u_1 \frac{\partial F}{\partial x} + u_3 \frac{\partial F}{\partial z}, \quad \frac{dz}{dt} = u_2 \frac{\partial F}{\partial x} - u_3 \frac{\partial F}{\partial y}$$

et le mouvement vers le maximum de  $z$  a lieu quand:

$$u_2 = a_1^2 \frac{\partial F}{\partial x}, \quad u_3 = -a_2^2 \frac{\partial F}{\partial y}$$

Le processus considéré peut être utilisé pour l'optimisation d'un paramètre composé dans les systèmes complexes, pour la construction du mouvement d'un robot et de systèmes ultra-stables.

## Zusammenfassung

Ein spezieller Zugang zu dem Problem der Suche nach den Extremwerten von Funktionen mit mehreren Veränderlichen an den Schnittstellen der mehrdimensionalen Flächen wird betrachtet. Die Differentialgleichungen, deren Lösungen auf diesen mehrdimensionalen Flächen (Mannigfaltigkeit) oder an ihren Schnittstellen festliegen, sind angegeben. Der Aufbau der Gleichungen enthält die willkürlichen Funktionen  $u_s$ , welche die Trajektorie auf den Flächen bestimmen. Im Falle der Darstellung der Mannigfaltigkeit  $F(x_1, x_2, \dots, x_n) = 0$  ist die Anzahl der willkürlichen Funktionen gleich  $C_n^2$ ; liegt jedoch die Trajektorie an den Schnittstellen von  $m$  Mannigfaltigkeiten  $F_j(x_1, x_2, \dots, x_n) = 0$  ( $j = 1, 2, \dots, m$ ), so ist die Anzahl der willkürlichen Funktionen  $C_n^{m+1}$ .

Hat man bestimmte konkrete willkürliche Funktionen  $u_s$ , so läßt sich eine Einrichtung finden, die kontinuierlich nach den Extremwerten der Funktionen mehrerer Variabler sucht; der Extremwert stellt den Endpunkt (eingeschwungener Zustand) der betrachteten Struktur dar. Durch die Wahl von verschiedenen Funktionen erreicht man, daß die Bewegungen zum Extremwert entlang verschiedenartiger Trajektorien (besonders entlang der geodätischen Linie) verlaufen. So sind zum Beispiel die Differentialgleichungen, deren Trajektorien auf der Fläche  $F(x, y, z) = 0$  liegen,

$$\frac{dx}{dt} = u_1 \frac{\partial F}{\partial y} - u_2 \frac{\partial F}{\partial z}, \quad \frac{dy}{dt} = -u_1 \frac{\partial F}{\partial x} + u_3 \frac{\partial F}{\partial z}, \quad \frac{dz}{dt} = u_2 \frac{\partial F}{\partial x} - u_3 \frac{\partial F}{\partial y}$$

Die Bewegung auf das Maximum von  $z$  hin findet statt, wenn

$$u_2 = a_1^2 \frac{\partial F}{\partial x}, \quad u_3 = -a_2^2 \frac{\partial F}{\partial y} \text{ ist.}$$

Die betrachtete Struktur kann zur Optimierung eines komplexen Systems, zur Lösung des Konstruktionsproblems einer Roboterbewegung und zur Erstellung eines ultrastabilen Systems dienen.

This paper considers a distinctive approach to the problem of synthesis of local systems for automatic search of extrema of functions of many variables. The principle involved in the construction of systems which react upon the partial derivatives of the sought function by the coordinates of the controlling devices is not new.

In the search of the extremum of the function of a single variable the problem is sufficiently defined; however, when the function depends upon several variables, the definiteness is lost and the solution of the problem becomes multi-valued. The function of a single variable  $y = f(x)$  is represented by a plane curve, and if at a certain point  $x_1$  one determines  $dy/dx$ , then it is necessary to vary  $x$  in order to approach the required extremum. A function of two variables may be represented by a surface  $y = f(x_1, x_2)$ . In this case, the path followed in passing

from a given point to the point of extremum, while remaining on the surface, is not a singular one, but one of infinite multitude.

In 1959 Krasovski considered systems which searched the path to the extremum by the gradient method<sup>7</sup>. He also showed that depending upon the form of the surface  $y = f(x_1, x_2, \dots, x_n)$  the gradient method may be varied by making the shift of the controlling device  $x_i$  dependent on the derivative  $f_i = \partial f / \partial x_i$  and also upon the derivatives of the function  $f$  with respect to other coordinates.

Approximately at the same time the Electro-Mechanical Institute of Leningrad considered the problem of simulating functions of many variables by means of digital differential analysers. This problem arose in connection with the construction of systems of programme control of metal cutting machines, first for simulating plane curves and then for curves lying on a given surface. The method which was utilized in this instance made it possible to indicate the general methods of synthesis with DDA (digital differential analyser) intended for simulating various forms of multidimensional surfaces, and also to indicate the quite general method of constructing systems for searching of extrema of functions of many variables, based on the principle of partial derivatives measurement. The gradient method is obtained in this instance as a special case. This method also permits the searching of extrema, taking into account the boundaries at the intersection of multidimensional surfaces.

Before proceeding to the treatment of this method, it is necessary to consider the question of the structure of differential equations whose solution lies at the given intersection of multidimensional surfaces<sup>4, 5</sup>.

#### The Structure of Differential Equations whose Solution Lies at the Intersection of Multidimensional Surfaces

The problem of finding differential equations whose solution is a given function is not a single-valued problem and its rational solution depends upon the means which are used for composing the sought system of equations, and upon which properties of the functions are utilized in the solution.

Indeed, in mathematical analysis there are given proofs of theorems on the singularity of solution of differential equations under given conditions; however, it is obvious that the converse problem has a multitude of solutions, that is, it is possible to find a multitude of differential equations whose singular solution will be the given function.

At the present time there exist two approaches for solving differential equations. The first approach permits the construction of an equation by introducing a parameter, and in the following this approach is called the parametric method of synthesis. The second approach of developing the method of synthesis is one which converts an equation into an identity, this equation being obtained by differentiating the output function with respect to the parameter.

Let there be a function

$$F(x_1, x_2, \dots, x_n) = 0 \quad (1)$$

and an argument  $\varphi$ . In the parametric method of synthesis one finds first of all the parametric equations

$$x_i = x_i(\varphi), \quad i = 1, 2, \dots, n \quad (2)$$

which satisfy (1), and from these equations differential equations are found whose solution will be given by functions (2). In this

case it is possible to find the differential equations by the method of  $K(D)$  transform proposed by Kulebakin.

By the second method<sup>4, 5</sup> the parametric expression (2) is not sought, but the differential equations whose solutions satisfy (1) are immediately determined. As numerous observations indicate, the structures synthesized by the second method are considerably simpler than the structures synthesized by the first one.

The basis of this method of analytical construction of a differential analyser is the following lemma: given the function (1) of  $n$  variables which has a derivative in the given range of the variables, then in order that the solution of differential equations

$$\frac{dx_i}{d\varphi} = f_i, \quad i = 1, 2, \dots, n \quad (3)$$

under initial conditions satisfying (1) may transform eqn (1) into an identity, it is necessary and sufficient that the eqns (3) transform into an identity eqn (4):

$$\sum_{i=1}^n \frac{\partial F}{\partial x_i} \frac{dx_i}{d\varphi} = 0 \quad (4)$$

This lemma is then utilized for searching functions  $f_i$ : they are sought so that they may transform eqn (4) into an identity. This problem has a multitude of solutions, and this is what determines the fact that the problem of synthesis is not a single-valued one. No matter how we may determine  $f_i$ , they will in all cases be some functions of partial derivatives  $\partial F / \partial x_i$ . In the operation the functions  $f_i$  are sought out as linear functions of partial derivatives under the assumption that this is the simplest case.

As is shown<sup>4, 5</sup> the differential eqns (3) whose solutions satisfy (1) include arbitrary functions  $U_s$  whose number is  $s = C_n^2$ , and the matrix of these arbitrary functions is symmetrical with respect to the diagonal with zeros along the principal diagonal. For instance, the structure of differential equations whose trajectories are disposed on surface  $F(x, y, z) = 0$  is determined by equations

$$\begin{aligned} \frac{dx}{d\varphi} &= u_1 \frac{\partial F}{\partial y} - u_2 \frac{\partial F}{\partial z} \\ \frac{dy}{d\varphi} &= u_1 \frac{\partial F}{\partial x} + u_3 \frac{\partial F}{\partial z} \\ \frac{dz}{d\varphi} &= u_2 \frac{\partial F}{\partial x} - u_3 \frac{\partial F}{\partial y} \end{aligned} \quad (5)$$

where  $u_1, u_2, u_3$  are arbitrary functions which determine the trajectory on the surface once they are given. They may be any functions as long as they satisfy Lipschitz conditions for right-hand sides of differential equations.

In simulating the trajectory at the intersection of surfaces

$$F_j(x_1, x_2, \dots, x_n) = 0 \quad j = 1, 2, \dots, m \quad (6)$$

$m < n$

the number of arbitrary functions  $u_s$  in the structure of differential equations is determined as

$$s = C_n^{m+1} \quad (7)$$

and it is possible to determine the disposition of these arbitrary functions in the structure of the equations.

As an illustration of these methods of synthesis differential equations will be found whose solutions are disposed on surfaces

$$F_j(x_1, x_2, x_3, x_4) = 0, \quad j = 1, 2 \quad (8)$$

At first the arbitrary functions are designated, the number of which in this case is  $C_4^3 = 4$ ,

$$u_1 = C_{123}, \quad u_2 = C_{124}, \quad u_3 = C_{134}, \quad u_4 = C_{234}$$

The coefficient in which the subscript of the term  $C$  contains unity are disposed in the first line; the coefficients in which this subscript contains the number two, are situated in the second line, etc., that is,

$$\begin{aligned} \frac{dx_1}{d\varphi} &= u_1 D_{23}^1 + u_2 D_{24}^1 + u_3 D_{34}^1 \\ \frac{dx_2}{d\varphi} &= -u_1 D_{13}^2 - u_2 D_{14}^2 + u_4 D_{34}^2 \\ \frac{dx_3}{d\varphi} &= u_1 D_{12}^3 - u_3 D_{14}^3 - u_4 D_{24}^3 \\ \frac{dx_4}{d\varphi} &= u_2 D_{12}^4 + u_3 D_{13}^4 + u_4 D_{23}^4 \end{aligned} \quad (9)$$

where the letter  $D$  designates the sum of the products of partial derivatives of the function (8) with respect to variables whose subscripts are present in the subscripts of the term  $D$ .

$$D_{12} = \frac{\partial F_1}{\partial x_1} \frac{\partial F_2}{\partial x_2} - \frac{\partial F_1}{\partial x_2} \frac{\partial F_2}{\partial x_1}, \quad \text{etc.}$$

The superscript of the term  $D$  denotes the line.

The signs before the terms in eqns (9), as can be shown, are determined by the following rule: consider the order of the superscript and subscript of the symbol  $D$ , for instance that normally indicated by a pointer, and if there is an odd number of violations of the normal order, a minus sign is used before this term, and in other cases a plus sign is used. That is, for terms of the top line one has orders 123, 124, 134 in which there are no violations, and these are accompanied by a plus sign; in the second line there are 213 with one violation (2 being greater than 1), and a minus sign is used; 214 with a minus sign, 234—no violations—a plus sign; in the third line, 312—two violations—plus sign; 314—one violation—minus sign; 324—one violation—minus sign; in the fourth line, 412—two violations—plus sign, and 413 and 423 also have plus signs.

In an analogous manner one determines the structure of differential equations and the signs and disposition of arbitrary functions in the latter by simulating trajectories at any number of intersecting surfaces with any number of variables.

It is of interest to note the presence of a maximum with respect to the number of arbitrary functions in the structure of differential equations for systems with a number of variables greater than six. The number of arbitrary functions for  $n < 6$  decreases as the number of intersecting surfaces increases. For  $n \geq 6$  the number of arbitrary functions for an increasing  $m$  at first increases, and only after having attained a maximum for  $m = (n - 2)$  with even values of  $n$ , and for  $m = (n - 2 \pm 1/2)$  for odd values of  $n$ , does it begin to decrease.

The arbitrary functions  $u_s$  in the structure of differential equations may be utilized as means of control in specifying prescribed motions on multidimensional surfaces, and as means

of self-tuning of an automatic control system. Formula (7) relates the number of dimensions of the control space to the number of degrees of freedom of an automatic control system, and to the number of constraints imposed upon the system, while the presence of a maximum in the number of arbitrary controlling functions indicates an optimal structure as regards self-tuning of a system for  $n \geq 6$ .

### Determination of Extrema of Functions

The problems of searching out the extrema of functions is one of the most widely encountered ones. There exist different methods of finding extrema in the presence of known partial derivatives, and different methods of automatic determination of these partial derivatives. However, at the present time, the methods of searching the extrema of functions in the presence of constraints placed upon the variables are not sufficiently well developed, and none of the existing methods assures that the motion to the extremum will proceed along a geodesic, or the shortest line.

In order to find the extrema one may utilize arbitrary coefficients in the structure of differential equations. Indeed, in order to assure the motion to an extremum—maximum with respect to coordinate  $x_i$ , it is sufficient to prescribe such a motion that the coordinate  $x_i$  increases all the time, and this may be achieved by specifying the coefficients  $u_s$  in a proper manner. For instance, in order to attain a maximum with respect to  $z$  on the surface  $F(x, y, z) = 0$  it is sufficient to assume in the system of eqns (5):

$$u_2 = a_1^2 \frac{\partial F}{\partial x}, \quad u_3 = -a_2^2 \frac{\partial F}{\partial y}$$

In this case

$$\begin{aligned} \frac{dx}{d\varphi} &= u_1 \frac{\partial F}{\partial x} - a_1^2 \frac{\partial F}{\partial x} \frac{\partial F}{\partial x} \\ \frac{dy}{d\varphi} &= -u_1 \frac{\partial F}{\partial x} - a_2^2 \frac{\partial F}{\partial y} \frac{\partial F}{\partial x} \\ \frac{dz}{d\varphi} &= a_1^2 \left( \frac{\partial F}{\partial x} \right)^2 + a_2^2 \left( \frac{\partial F}{\partial y} \right)^2 \end{aligned} \quad (10)$$

where  $dz/d\varphi$  will be a positive definite form of a constant sign for all real values of  $x, y, z$ , which assures the stability of the process of finding the extremum in accordance with Liapunov<sup>6</sup>. At the point of the maximum with respect to  $z$ , the velocities with respect to all coordinates become zero. For a system of eqns (10) the point of maximum with respect to  $z$  proves to be a point of stable equilibrium. In the motion toward the extremum—minimum

$$u_2 = -a_1^2 \frac{\partial F}{\partial x}, \quad u_3 = a_2^2 \frac{\partial F}{\partial y}$$

and  $dz/d\varphi$  will be a negative definite form.

In an analogous manner we determine the coefficients  $u_s$  for a specified motion toward the extremum for surfaces with a large number of dimensions as well.

The synthesized structures may be utilized for searching out extrema of functions with any number of variables for individual surfaces as well as for cases in which constraints are taken into account, that is, for intersecting surfaces. For instance, in searching the maximum with respect to coordinate ( $x_4$ ) at

the intersection of surfaces (8) for a specified motion toward this extremum it is possible in the system of eqns (9) to let

$$u_2 = D_{12}, u_3 = D_{13}, u_4 = D_{23}$$

and  $dx_4/d\varphi$  will be a positive definite form, and this fact assures stability of the process of searching the extremum according to Liapunov.

For a motion toward the extremum prescribed in this manner there remain free arbitrary functions in the synthesized structures, the number of which functions is equal to:

$$s = C_n^{m-1} - C_{n-1}^m$$

These free arbitrary functions may be utilized for correcting the trajectory during the time of the motion toward the extremum. In the example considered above there remains a free arbitrary function  $u_1$  in the system of eqns (10). The free arbitrary functions may be utilized for prescribing the motion toward the extremum along a geodesic curve or one which is close to it.

It should be noted that all stationary points for the obtained differential equations will be points of equilibrium, but only points of the extrema will be points of stable equilibrium, while the saddle points will be points of unstable equilibrium.

If the number of intersecting manifolds is  $m = n - 1$ , then they determine a line in the  $n$ -dimensional space. In this case the problem is reduced to searching out the extremum in a one-dimensional manifold. It may be assumed that  $d\varphi = \omega dt$ , where  $t$  is the time and  $\omega$  is an arbitrary function which satisfies the Lipschitz condition, and

$$\frac{dx_i}{dt} = \omega \xi_i(x_1, x_2, \dots, x_n), \quad i = 1, 2, \dots, n$$

For prescribing the motion toward the extremum in this case it is only necessary to specify the direction of the motion along the line. For example, in the motion toward the maximum with respect to  $x_n$  it is sufficient to assume that  $\omega = \xi_n(x_1, x_2, \dots, x_n)$ , and then

$$\frac{dx_i}{dt} = \xi_n(x_1, x_2, \dots, x_n) \cdot \xi_i(x_1, x_2, \dots, x_n)$$

$$\frac{dx_n}{dt} = \xi_n^2(x_1, x_2, \dots, x_n)$$

There follows a comparison of the described method of searching out the extrema and the gradient method. As shown by Krasovski<sup>7</sup>, the gradient method assures stability, according to Liapunov, in the computing process of searching the extremum. This constitutes the similarity between them. But the gradient method assures the displacement toward the extremum only along some special trajectory, while the proposed method permits the variation of trajectory of motion toward the extremum.

Indeed, in the system of eqns (10) there remained one free arbitrary coefficient  $u$  which may be specified by a different method and which supplements the definition of trajectory for the motion toward the extremum. An analogous situation exists also in searching the extremum for other manifolds or their intersections, except for those which are one-dimensional. The gradient method constitutes a special case of the considered method of searching the extrema, when all the remaining arbitrary coefficients are set equal to zero; for instance, for the system of eqns (10), when  $u_1 = 0$ .

The remaining arbitrary coefficients may be prescribed in

such a manner as to assure the motion toward the extremum along a trajectory which is optimal in some sense, including in this number a geodesic trajectory.

Figure 1 shows a block diagram of a system which searches out an extremum at the intersection of surfaces. The controlling signals produced by an analogue programming device (PD) are supplied to several simultaneously optimized plants  $O_1, O_2, \dots, O_m$ . In these plants the current values of partial derivatives which are supplied to the programming device are determined in some manner. The programming device constitutes a differential analyser (in particular, an electronic analogue installation) whose structure was described in the preceding paragraph. The setter of trajectories (ST) carries out such prescriptions of the arbitrary coefficients which remain free after the prescription of motion toward the extremum in order to assure the displacement toward it along some desired trajectory.

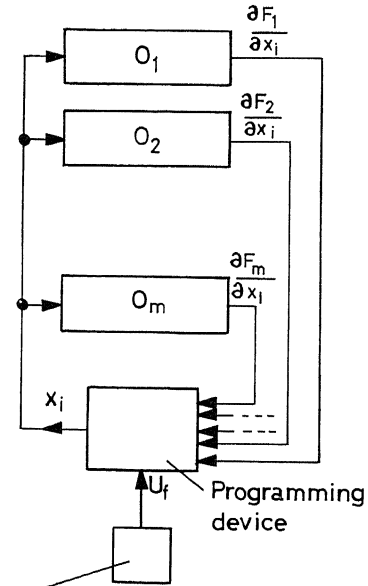


Figure 1

If the equations  $F_i(x_1, x_2, \dots, x_n) = 0, j = 1, 2, \dots, m$  are known, then  $O_1, O_2, \dots, O_m$  are simply functional transforms. If only a part of these equations is known, this means that a part of  $O_1, O_2, \dots, O_m$  are functional transforms (computer assemblies), while the other part are the plants.

In Figure 2 is shown the block diagram of a system which utilizes the method of searching the extremum described above. As an example consider the case of searching the maximum on a surface  $F(x_1, x_2, x_3, x_4) = 0$  with respect to coordinate  $x_4$ . The structure of the analogue device in this case is defined by equations

$$\begin{aligned} \frac{dx_1}{d\varphi} &= u_1 \frac{\partial F}{\partial x_2} - u_2 \frac{\partial F}{\partial x_3} - \frac{\partial F}{\partial x_1} \frac{\partial F}{\partial x_y} \\ \frac{dx_2}{d\varphi} &= -u_1 \frac{\partial F}{\partial x_1} + u_4 \frac{\partial F}{\partial x_3} - \frac{\partial F}{\partial x_2} \frac{\partial F}{\partial x_y} \\ \frac{\partial x_3}{\partial \varphi} &= u_2 \frac{\partial F}{\partial x_1} - u_4 \frac{\partial F}{\partial x_2} - \frac{\partial F}{\partial x_3} \frac{\partial F}{\partial x_y} \\ \frac{dx_y}{d\varphi} &= \left( \frac{\partial F}{\partial x_1} \right)^2 + \left( \frac{\partial F}{\partial x_2} \right)^2 + \left( \frac{\partial F}{\partial x_3} \right)^2 \end{aligned} \quad (11)$$

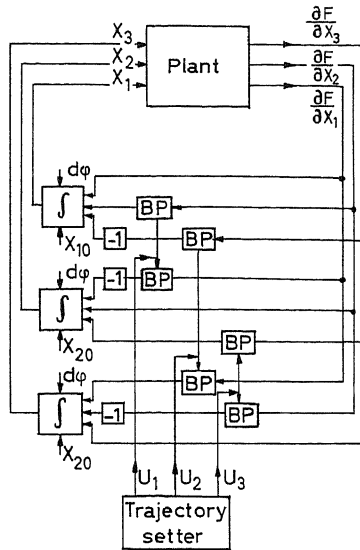


Figure 2

In this instance we assume that  $\partial F / \partial x_4 = -1$ . The current values of partial derivatives may be determined by the method of synchronous detection. The considered system for  $u_1 = u_2 = u_3 = 0$  is transformed into a scheme of extremal system cited by Krasovski<sup>7</sup> and it differs from this scheme by the introduction of cross-links supplied to the input of the integrators. At the same time, the coefficients  $u_1, u_2, u_3$  may be either constant magnitudes or functions of coordinates  $x_i$ , and be controlled by some index of the quality of operation of the system.

In specifying the motion along a geodesic curve in eqn (10), the free coefficient  $u_1$ , for instance, may be determined from the condition that for a geodesic curve the main normal to the curve coincides with the normal to the surface, and at the same time  $u_1$  is determined as a complex function of coordinates.

If we search an extremum with respect to coordinate  $y$  on the surface  $F(x, y, t) = 0$ , where  $t$  is the time, then the structure of the analyser which specifies the motion toward the extremum will be defined by equations

$$\begin{aligned} \frac{dx}{dt} &= -\frac{\partial F}{\partial x} \frac{\partial F}{\partial y} - u_2 \frac{\partial F}{\partial t} \\ \frac{dy}{dt} &= \left( \frac{\partial F}{\partial x} \right)^2 + \left( \frac{\partial F}{\partial t} \right)^2 \\ 1 &= u_2 \frac{\partial F}{\partial x} - \frac{\partial F}{\partial t} \frac{\partial F}{\partial y} \end{aligned}$$

As can be seen, by virtue of the last equation of this system of equations, the number of free arbitrary coefficients decreases.

It is possible to determine such constant coefficients  $u_s$  which assure the motion toward the extremum, perhaps not along the geodesic curve but at least along a path which is shorter than the trajectories followed during the motion toward the extremum by the gradient method, that is, when the free arbitrary coefficients are equal to zero. During the motion along a geodesic curve these coefficients in the general case will be complex functions. For constant free arbitrary coefficients, the technical realization of the proposed method is considerably simplified.

## On the Search of Extrema of Functions in Automatic Control Systems

The operation involved in searching out extrema at the present time is automated to a large extent and may be used as a basis of construction of various automatic control systems. In the case of a limited range of change of variables the extremum may be sought taking into account the constraint.

$$\sum x_i^2 = R^2$$

The method described above permits this approach. Frequently in controlling chemical production of great complexity the problem of optimization of the free index of the quality of the process arises; for instance, if there is a plant with a characteristic  $F(x, y, z) = 0$ , and it is required to determine such values of  $x, y, z$  which would provide an extremum to the free index  $z' = l_3 x + m_3 y + n_3 z$  where  $l_3, m_3, n_3$  are constant quantities, then this problem may also be solved on the basis of the method considered above.

Rewriting these equations using other designations, one has

$$F_1(x_1, x_2, x_3) = 0$$

$$F_2 = l_3 x_1 + m_3 x_2 + n_3 x_3 - x_4 = 0$$

The structure of differential equations whose solution lies at the intersection of these surfaces is determined as (9), where

$$D_{12} = m_3 \frac{\partial F_1}{\partial x_1} - l_3 \frac{\partial F_1}{\partial x_2}$$

$$D_{13} = n_3 \frac{\partial F_1}{\partial x_1} - l_3 \frac{\partial F_1}{\partial x_3}$$

$$D_{1y} = -\frac{\partial F_1}{\partial x_1}, D_{23} = n_3 \frac{\partial F_1}{\partial x_2} - m_3 \frac{\partial F_1}{\partial x_3}$$

$$D_{2y} = -\frac{\partial F_1}{\partial x_2}, D_{3y} = -\frac{\partial F_1}{\partial x_3}$$

The partial derivatives of the characteristic of the plant may be determined by some automatic method<sup>1, 2</sup>.

The problem considered above may be formulated as a problem of searching an extremum in a given direction, which is characterized by coefficients  $l_3, m_3, n_3$ . At the present time an effort is being made to utilize the operation of searching extrema for solving the problem of constructing the motions<sup>10</sup>. The problem of constructing the motions based on synergy levels<sup>11, 12</sup> may be formulated for the given kinematic scheme in terms of the intersections of the manifolds, and the motions themselves may be regarded as a solution of the problem of searching an extremum in a given direction.

In conclusion, consider the problem of possibilities of a global search. The finding of an extremal extremum requires a more thorough study of the investigated functions, and at the present time various strategies for solving this problem<sup>1, 10, 13</sup> have been proposed. One can propose yet another strategy for solving this problem as follows. Suppose that it is necessary to find the maximal maximum. Having investigated the function and having found several maxima, it is possible to pass a surface through them and the maximum of this surface will be

at least in the zone of gravity of the sought maximum of the maxima. If the approximated surface will have several maxima, then it may be smoothed in the same manner by finding the second approximated surface, etc. The number of approximated surfaces will be determined by the complexity of the investigated function.

## References

- <sup>1</sup> FELDBAUM, A. A. *Computers in Automatic Control Systems*. 1959. Moscow; Fizmatgiz
- <sup>2</sup> IVANOV, V. N. On the determination of partial derivatives of functions of many variables in systems of automatic control. *Izvestiya AN SSSR, OTN, Energetika i avtomatika* No. 4 (1960)
- <sup>3</sup> HOERL, A. E. *A Technique for Optimizing Process Conditions*. Fourth ASME/I.R.D. Conference, Newark, Delaware, April 1958
- <sup>4</sup> IGNATJEV, M. B. Synthesis of differential analysers for reproducing implicit functions. Collected works on problems of electro-mechanics. *Izd. AN SSSR* No. 5 (1961)
- <sup>5</sup> IGNATJEV, M. B. On the problem of synthesis of differential analysers. *Izvestiya AN SSSR, OTN, Energetika i avtomatika* No. 2 (1961)
- <sup>6</sup> LIAPUNOV, A. M. *General Problem of Stability of Motion*. 1935. ONTI
- <sup>7</sup> KRASOVSKI, A. A. Dynamics of continuous systems of extrema regulation based on the gradient method. *Izvestiya AN SSSR, OTN, Energetika i avtomatika* No. 3 (1959)
- <sup>8</sup> IGNATJEV, M. B. On the searching of extrema of functions with the aid of electronic analogs. *Reports of 4th Interuniversity Conference on the Application of Analogs in Various Branches of Technology*. Vol. 3, 1962. Izd. MEI
- <sup>9</sup> IGNATJEV, M. B. *Certain Problems of Synthesis and Application of Differential Analysers as Control Devices*. Author's synopsis of dissertation. (1962)
- <sup>10</sup> GELFOND, Tz. M., TZETLIN, M. L. On several methods of control of complex systems. *Progress of Mathematical Sciences (Uspekhi matematicheskikh nauk)*. XVII, No. I (1962) 103
- <sup>11</sup> GAMBARYAN, L. S. *Problems of Physiology of a Motion Analyzer*. 1962. Medgiz
- <sup>12</sup> BERNSTEYN, N. A. *On Construction of Motions*. 1947. Medgiz
- <sup>13</sup> BOCHAROV, I. N., FELDBAUM, A. A. An automatic optimizer for searching the minimal of several minima. *Automat. telemek.* No. 3 (1962)
- <sup>14</sup> VORONOV, A. A. *et al.* Digital analogs for systems of automatic control. *Izd. AN SSSR, M. — L.* (1960)

## DISCUSSION

R. KULIKOWSKI, *Polish Academy of Sciences, Warsaw, Poland*

The random behaviour of the plant characteristics is a distinctive feature of the problems which are connected with extrema searching in automatic control systems. Therefore it is very important to minimize the searching time of the risk function associated with the given performance measure and plant characteristics. The performance measure is usually a functional of the time-dependent control signals rather than a function of the static input values. Consequently, the generalized gradient which determines the changes of the control signals is a time-dependent operator. In order to find the extremum by the steepest descent, or other similar method, we have to identify the generalized gradient at every step of iterations. All these factors complicate the problem under consideration. However, some important questions connected with the minimization of the risk function were already considered by A. A. Feldbaum in his papers on the so-called dual control. The difficulties connected with the identification of the generalized gradient are considered in my paper presented at this congress.

In view of these difficulties it would be interesting to know what kind of practical problems can be treated by the method described in the very interesting paper of A. A. Voronov and M. B. Ignatjev, and

how the system operates when the memory and random changes of the plant characteristics cannot be ignored?

A. A. VORONOV, *in reply*

Two important problems not considered in the report are: (1) dynamical properties of the controlled system and (2) the influence of random noise on the searching of extremum.

Dynamics, in general, may be taken into account by the addition of the dynamic equation to the ones considered; this equation may be treated as one of the equations of surfaces (or manifolds). Thus the method may be applied in this case also.

The influence of random noise demands a special investigation. This question is very important, yet the subject of the report was not the dynamics of searching but only the principle of the device which performs the moving to the extremum of the set function.

The minimization of searching time is a very important problem. We may minimize the time if we choose the correct method of searching. The correct method, however, depends on the form of function  $F(x_1, \dots, x_n)$ . In some cases the device considered may be useful for discovering the correct method of searching.



# Dominant Operators Approach to the Theory of Adaptive Control Systems

A. STRASZAK

## Summary

This paper considers an approach to the synthesis of an adaptive control system that provides for rapid adaptations. It leads to a model adaptive control system, but not to a model comparison system. A model is used as a dominant operator, and then the dynamics of process may be neglected. Time-domain synthesis of dominant operator is introduced. In the time-domain synthesis, performance index approach and model comparison approach may be transferred to the desired form of differential equation of error signal. Advantages of this type of adaptive control system are studied. An example and some experimental results are given.

## Sommaire

Ce papier présente une méthode pour la synthèse d'un système de commande adaptative permettant des adaptations rapides. Le système comprend un modèle qui n'est pas utilisé comme comparateur, mais comme 'opérateur dominant'. De ce fait, la dynamique du processus commandé peut être négligée. On introduit la synthèse de l'opérateur dominant dans le domaine du temps. Dans la synthèse, effectuée dans le domaine du temps, la méthode de l'index de performance et celle de la comparaison avec le modèle peuvent être ramenées à la forme voulue d'une équation différentielle avec le signal d'erreur comme variable. On étudie les avantages de ce type de systèmes de commande adaptative. On donne un exemple et des résultats expérimentaux.

## Zusammenfassung

Der Aufsatz behandelt ein Verfahren zur Synthese von selbsteinstellenden Regelsystemen, das eine schnelle Selbsteinstellung ermöglicht. Dies führt zu einem Modell eines selbsteinstellenden Regelsystems; es handelt sich aber nicht um einen Vergleich mit einem Bezugsmodell (Modellvergleich). Das Modell dient vielmehr als dominanter Operator (der das dynamische Verhalten des Systems beschreibt), daher läßt sich das dynamische Verhalten der Regelstrecke vernachlässigen. Es wird die Synthese des dominierenden Operators im Zeitbereich eingeführt. Bei der Synthese im Zeitbereich lassen sich das Verfahren des Gütekriteriums und der Modellvergleich in die gewünschte Form der Differentialgleichung für die Regelabweichung überführen. Die Arbeit untersucht die Vorteile dieser Art der selbsteinstellenden Regelung. Ein Beispiel und einige Versuchsergebnisse sind angegeben.

## Introduction

Several different philosophical approaches to defining satisfactory performance of an adaptive control system have been offered; two of the most common are the performance criterion, and model comparison. In performance criterion approach some generalized performance index, such as the integral of squared error, is chosen and is computed continuously.

This information is used to adjust the proper system parameters in the required manner to keep the value of the index at a minimum. In model comparison approach a specific model is

chosen which represents the desired system characteristics. If the parameters of the system deviate from those specified by the model, they are suitably adjusted. These two general procedures are illustrated in *Figure 1*.

The principal drawback to the first scheme has been the inability to arrive at a performance index which will provide a satisfactory system characteristic for all inputs. Another disadvantage to the first scheme has been the delay in the performance index control loop.

The second scheme also, however, presents a fundamental obstacle, in that the procedures required to calculate the existing system parameters usually involve correlation techniques which require integration and averaging of input and output signals over periods of time that are long compared with system time constants.

Hence in these situations the process parameters are constrained to vary slowly compared with either the system dynamics or the input signal, in order that the parameters can be identified

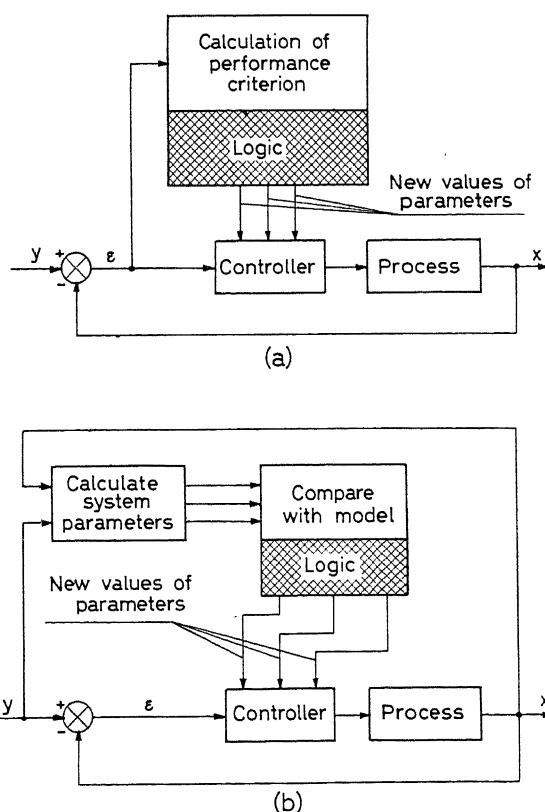


Figure 1. Two classes of adaptive controller: (a) performance index; (b) model comparison

before they change appreciably. Often this constraint is easily met—as in the case of many industrial adaptive control requirements—but rapid systems are necessary for aircraft and space vehicles.

### Dominant Operators

Ten years ago, Chu<sup>2</sup> proposed synthesis of linear control system by predominant roots. In this paper it is assumed that all the modes of signal output equation may be neglected except the first constant term, due to the pair of predominant roots. This approach was used to the synthesis of adaptive system by Staffin<sup>9</sup>, who has proposed a system which provides for neglecting process roots. As shown in Figure 2 this is a model system but not a model comparison system. A model is used in this system as a dominant operator and then dynamics of process may be neglected. This approach was used also by Li<sup>4</sup>, Mellen<sup>6</sup>, and Schuck<sup>13</sup>.

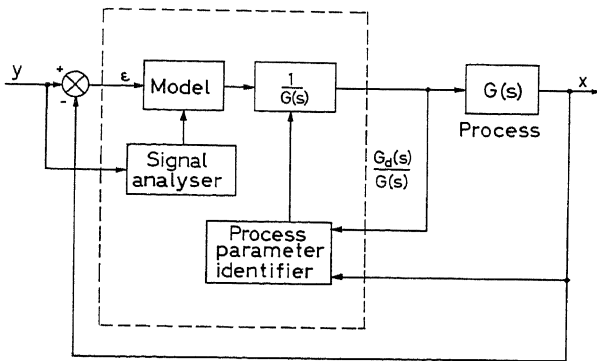


Figure 2. Staffin's executive-controlled adaptive system

Staffin has proposed  $s$ -domain synthesis of a dominant operator adaptive system, but this method is useless for rapid adaptations because identification of process parameters is needed.

A more useful system for rapid adaptations may be obtained when a time-domain synthesis of a dominant operator is used. This method has been based on the earlier works of Straszak<sup>10,11</sup>.

### Time-Domain Synthesis of the Dominant Operator

To keep the value of the performance index at a minimum or to obtain the desired system characteristic is the goal of the adaptive control system.

Consider first the minimum performance index system. Usually, in control systems the performance index is a function of the error signal<sup>14</sup>. Thus, the performance index,  $K$ , is

$$K = \int_{t_2}^{t_1} F[\varepsilon(t), \dot{\varepsilon}(t), \dots, \varepsilon^{(k)}(t), t] dt \quad (1)$$

Now, it is necessary to find an operator (dynamic system) which will minimize the performance index  $K$ .

If it is assumed that a function  $F$  is known,  $k+2$  times the differential function of all the arguments, then very well-known variational methods may be used to obtain the Euler-Poisson equation in the following form.

$$F_\varepsilon - \frac{d}{dt} F_{\varepsilon'} + \frac{d^2}{dt^2} F_{\varepsilon''} + \dots + (-1)^k \frac{d^k}{dt^k} F_{\varepsilon^{(k)}} = 0 \quad (2)$$

where

$$F_\varepsilon = \frac{\partial F}{\partial \varepsilon}$$

$$F_{\varepsilon'} = \frac{\partial F}{\partial \varepsilon'} = \frac{\partial F}{\partial \left( \frac{d\varepsilon}{dt} \right)}$$

$$\vdots$$

$$F_{\varepsilon^{(k)}} = \frac{\partial F}{\partial \varepsilon^{(k)}} = \frac{\partial F}{\partial \left( \frac{d^k \varepsilon}{dt^k} \right)}$$

The general solution of eqn (2) has  $2k$  arbitrary constants, which may be found from the boundary points

$$\varepsilon(t_0) = \varepsilon_0, \varepsilon'(t_0) = \varepsilon'_0, \dots, \varepsilon^{(k-1)}(t_0) = \varepsilon_0^{(k-1)}$$

$$\varepsilon(t_1) = \varepsilon_1, \varepsilon'(t_1) = \varepsilon'_1, \dots, \varepsilon^{(k-1)}(t_1) = \varepsilon_1^{(k-1)}$$

Thus, when eqn (2) is solved the following is obtained:

$$\varepsilon(t) = f[t, \varepsilon_0, \varepsilon'_0, \dots, \varepsilon_0^{(k-1)}, \varepsilon_1, \varepsilon'_1, \dots, \varepsilon_1^{(k-1)}] \quad (3)$$

In many cases it can be assumed that

$$\varepsilon_1 = 0, \varepsilon'_1 = 0, \dots, \varepsilon_1^{(k-1)} = 0 \quad (4)$$

Putting (4) into (3), the following form of eqn (3) is obtained:

$$\varepsilon(t) = f^*[t, \varepsilon(t_0), \dot{\varepsilon}(t_0), \dots, \varepsilon(t_0)^{(k-1)}] \quad (5)$$

Differentiating (5), the following is obtained:

$$\frac{d\varepsilon(t)}{dt} = g[t, \varepsilon(t_0), \dot{\varepsilon}(t_0), \dots, \varepsilon(t_0)^{(k-1)}] \quad (6)$$

Applying the principle of optimality<sup>1</sup> it can be assumed that

$$t \equiv t_0 \equiv 0 \quad (7)$$

By putting the problem (7) into form (6), a differential equation is obtained:

$$\frac{d\varepsilon}{dt} = h(t, \varepsilon, \dots, \varepsilon^{(k-1)}) \quad (8)$$

of dynamic system (operator) which will be minimized by the performance index  $K(\varepsilon)$

The problem becomes more complicated when the performance index  $K$  needs to be minimized, subject to the constraints

$$\phi(t, \varepsilon, \dot{\varepsilon}, \dots, \varepsilon^{(k)}) = 0$$

or

$$\int_{t_2}^{t_1} F_i(t, \varepsilon, \dot{\varepsilon}, \dots, \varepsilon^{(k)}) dt = l_i$$

but this also may be solved by using Lagrange's method of multipliers. When classical variational methods cannot be used,

the methods of Pontriagin, Bellman or Kulikowski<sup>3</sup> may be applied.

When a solution of a variational problem is to be obtained a procedure from (3) ÷ (8) may be used and then a desired differential equation of a dynamic system which will minimize the performance index  $K(\varepsilon)$  subject to the constraints is obtained.

In model comparison approach a specific model is chosen, which represents the desired system characteristics. Usually, a model has been known in transfer function form, but it is very easy to transform a desired system characteristic to the differential equation form, as (8).

Thus, from the time-domain synthesis approach, the performance index approach and model comparison may be transformed into the one form, desired differential equation of error signal.

Now in a general case it can be assumed that the desired differential equation in form (8) is given.

Consider accordingly the adaptive control system with the following desired differential equation of error signal.

$$\frac{d\varepsilon}{dt} = r(\varepsilon, \dots, \varepsilon, t) \quad (9)$$

The form of the physical adaptive control system chosen for time-domain synthesis is shown in Figure 3. Although this

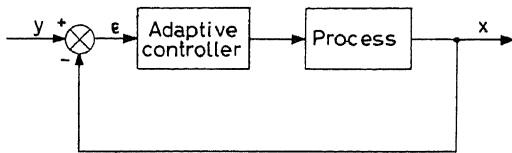


Figure 3. Adaptive control system chosen for time domain synthesis

form is not essential for the application of the synthesis procedure, the explanation of the operation of the system is greatly facilitated, especially for the dominant operator-adaptive control system.

Assume that the process may be described by a differential equation which is not known,

$$\frac{dy}{dt} = p(y, \dots, y, u, t) \quad (10)$$

but a solution of this equation may be observed. Now we can formulate our goal: we want to find a control signal  $u(t)$  which will give desired system characteristic.

Consider now the error signal in this system. It may be assumed that the control signal  $u(t)$  is not changed for a fixed period of time  $t_0 \leq t \leq t_1$  and observe the error signal as a function of time (Figure 4). On the same figure we may draw the solution of the desired differential equation with the same initial values in point  $t_0$ .

Dividing by  $\tau$  time axis, two sequences are obtained:

$$\{\varepsilon_1(t_0), \varepsilon_1(t_0 + \tau), \varepsilon_1(t_0 + 2\tau), \dots, \varepsilon_1(t_1)\}$$

and

$$\{\varepsilon_2(t_0), \varepsilon_2(t_0 + \tau), \varepsilon_2(t_0 + 2\tau), \dots, \varepsilon_2(t_1)\}$$

where

$\varepsilon_1$  is the desired value, and

$\varepsilon_2$  the uncontrollable value.

Suppose that system has been at  $t_0 + \tau$  point.

At this point, there have been two values  $\varepsilon_2(t_0 + \tau)$  and  $\varepsilon_2(t_0 + \tau)$ . To obtain the desired value only the control value  $u$  can be changed.

This value may be changed at random because it is not known how much change is necessary. Thus an additional function is obtained:  $\varepsilon_{\Delta u}(t)$  (Figure 4).

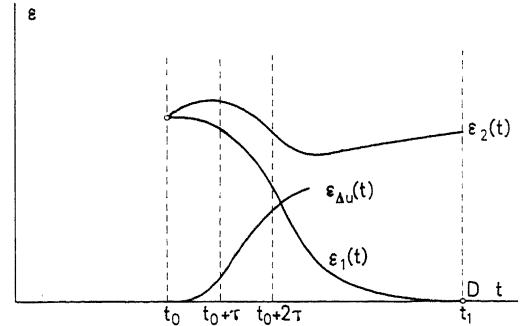


Figure 4. Errors signals:  $\varepsilon_1(t)$ —desired function;  $\varepsilon_2(t)$ —uncontrollable signal;  $\varepsilon_{\Delta u}(t)$ —error signal response for  $\Delta u_1$

Usually, for most of the processes,  $\varepsilon_{\Delta u}(t)$  is equal to zero at the starting point and only this type of process is considered.

Therefore, to obtain the desired value  $\varepsilon_1(t_0 + \tau)$  the control signal  $u(t)$  must be changed some time before; for example at  $t_0$ .

Error signal  $\varepsilon(t_0 + \tau)$  may be written in the following form.

$$\begin{aligned} \varepsilon(t_0 + \tau) &= x(t_0 + \tau) - y(t_0 + \tau) = x(t_0) - y(t_0) \\ &+ \left( \frac{dx}{dt} - \frac{dy}{dt} \right) \tau + \int_{t_0}^{t_0 + \tau} h_u(\tau - x) du(x) \\ &+ \int_{t_0}^{t_0 + \tau} h_b(\tau - x) db(x) \pm y + R_x + R_y \dots \end{aligned} \quad (11)$$

where

$h_u, h_b$  is the response of system to unit step control signal  $u$  and disturbance signal  $b$ ,

$\Delta y$  is the jumping function, and

$R_x, R_y$  the second order small value.

In eqn (11) there is only one controllable term, as

$$\int_{t_0}^{t_0 + \tau} h_u(\tau - x) du(x)$$

Now,  $\tau$  will be very small, so it may be assumed that

$$\int_{t_0}^{t_0 + \tau} h_b(\tau - x) db(x) \approx 0$$

and

$$R_x, R_y \approx 0$$

Jumping function  $\Delta y$  may be removed to the next point.

Therefore, the necessary increment of control value at  $t_0$ , for obtaining the desired error value at  $t_0 + \tau$ , is<sup>11</sup>

$$\Delta u = \frac{\varepsilon_1(t + \tau) - \varepsilon_2(t + \tau)}{h_u(\tau)} \quad (12)$$

This may be rewritten as follows:

$$\Delta u(t) = \frac{\tau}{h_u(\tau)} \left[ \frac{\varepsilon_1(t)}{\tau} + \frac{d\varepsilon_1(t)}{dt} - \frac{\varepsilon_2(t)}{2} - \frac{d\varepsilon_2(t)}{dt} \right] \quad (13)$$

since

$$\varepsilon_1(t) \equiv \varepsilon_2(t)$$

$$\Delta u(t) = k \left[ \frac{d\varepsilon_1(t)}{dt} - \frac{d\varepsilon_2(t)}{dt} \right] \quad (14)$$

where

$$k = \frac{\tau}{h_u(\tau)}$$

Now, substituting (9) into (14), gives

$$\Delta u(t) = k \left[ r(\varepsilon, \dots, \varepsilon, t) - \hat{\varepsilon} \right] \quad (15)$$

$k$  is unknown *a priori*, but assuming this value at random is not necessary from the practical point of view.

Usually, it is not very difficult first to estimate maximum and minimum values of  $h_u(\tau)$  and then to obtain average value of  $h_u(\tau)$  as follows

$$h_{u\text{avr}}(\tau) = \frac{h_u(\tau)_{\max} + h_u(\tau)_{\min}}{2} \quad (16)$$

$\tau$  must be chosen much smaller than time constants of the process. The physical realization of this procedure is illustrated in Figure 5.

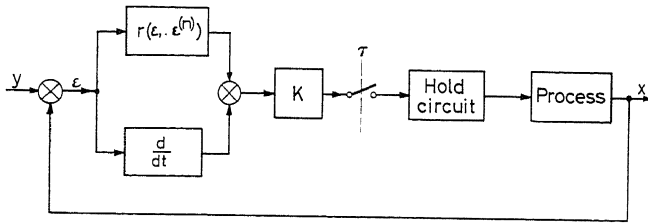


Figure 5. Dominant operator adaptive control system

Now, we consider a sensitivity of inaccuracy of gain coefficient  $k$  on the system characteristic.

Now, let  $\delta\varepsilon = \varepsilon_1(t+\tau) - \varepsilon(t+\tau)$  be an adaptive error function and

$$\rho = \frac{kh_u(\tau)}{\tau}$$

be an inaccuracy of gain coefficient.

Then

$$\begin{aligned} \delta\varepsilon^I &= \varepsilon_1(t+\tau) - \varepsilon(t+\tau) = \left[ (1-\rho) \left( \frac{d\varepsilon_1}{dt} \right)_t + (\rho+1) \left( \frac{d\varepsilon}{dt} \right)_t \right] \tau \\ &= (\rho-1) \left[ \left( \frac{d\varepsilon}{dt} \right)_t - \left( \frac{d\varepsilon_1}{dt} \right)_t \right] \tau \end{aligned}$$

$$\begin{aligned} \delta\varepsilon^{II} &= \varepsilon_1(t+2\tau) - \varepsilon(t+2\tau) = (\rho-1) \left[ \left( \frac{d\varepsilon}{dt} \right)_{t+\tau} - \left( \frac{d\varepsilon_1}{dt} \right)_{t+\tau} \right] \tau \\ \left( \frac{d\varepsilon}{dt} \right)_{t+\tau} &= \left( \frac{d\varepsilon}{dt} \right)_t + \rho \left[ \left( \frac{d\varepsilon_1}{dt} \right)_t - \left( \frac{d\varepsilon}{dt} \right)_t \right] \end{aligned}$$

$$\delta\varepsilon^{II} = (\rho-1) \left[ \left( \frac{d\varepsilon}{dt} \right)_t + \rho \left\{ \left( \frac{d\varepsilon_1}{dt} \right)_t - \left( \frac{d\varepsilon}{dt} \right)_t - \left( \frac{d\varepsilon_1}{dt} \right)_{t+\tau} \right\} \right] \tau$$

$$= (\rho-1) \left[ \left( \frac{d\varepsilon}{dt} \right) (1-\rho) + \rho \left\{ \left( \frac{d\varepsilon_1}{dt} \right)_t - \left( \frac{d\varepsilon_1}{dt} \right)_{t+\tau} \right\} \right] \tau$$

$$= (\rho-1) \left[ \left( \frac{d\varepsilon}{dt} \right)_t (1-\rho) - \left( \frac{d\varepsilon_1}{dt} \right)_t (\eta-\rho) \right] 2$$

where

$$\eta = \left( \frac{d\varepsilon_1}{dt} \right)_{t+\tau} : \left( \frac{d\varepsilon_1}{dt} \right)_t$$

Assume

$$\eta \approx 1$$

then

$$\delta\varepsilon^{II} = \left\{ (\rho-1)(1-\rho) \left[ \left( \frac{d\varepsilon}{dt} \right)_t - \left( \frac{d\varepsilon_1}{dt} \right)_t \right] \right\} \tau$$

For

$$|\rho| < 2$$

$$|\delta\varepsilon^{II}| < |\delta\varepsilon^I|$$

$$\text{sign } \delta\varepsilon^{II} = -\text{sign } \delta\varepsilon^I$$

Now, it can be concluded that this system has self-correcting ability; this property was confirmed by experimental study (see example).

If a dynamic characteristic of process is changing too much an additional control loop may be used to endeavour to solve the following equation

$$\frac{h(\tau)}{\tau} = k = \text{const} \quad (17)$$

by adjusting the sampling interval  $\tau$ . Multiplying (17) by  $\Delta u(t+\tau)$  and  $\tau$

$$\Delta u(t+\tau) h(\tau) = k \cdot \tau \cdot \Delta u(t+\tau) \quad (18)$$

is obtained since

$$x(t+\tau) - x(t) - \frac{dx}{dt} \cdot \tau \approx \Delta u(t+\tau) h(\tau) \quad (19)$$

Substituting (19) into (18), gives

$$x(t+\tau) - x(t) - \left( \frac{dx}{dt} \right) \tau = k \cdot \tau \cdot \Delta u(t+\tau) \quad (20)$$

This equation may be solved automatically as shown in Figure 6.

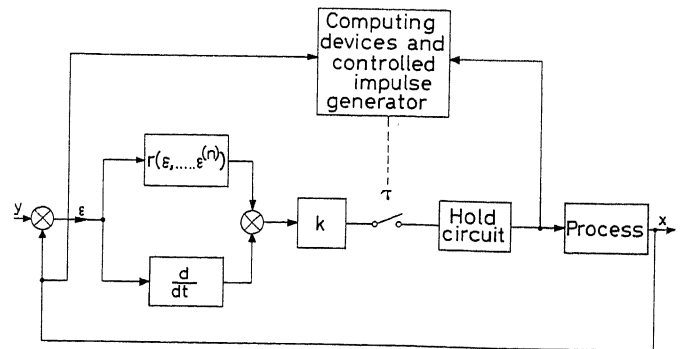


Figure 6. Dominant operator adaptive control system with additional control loop

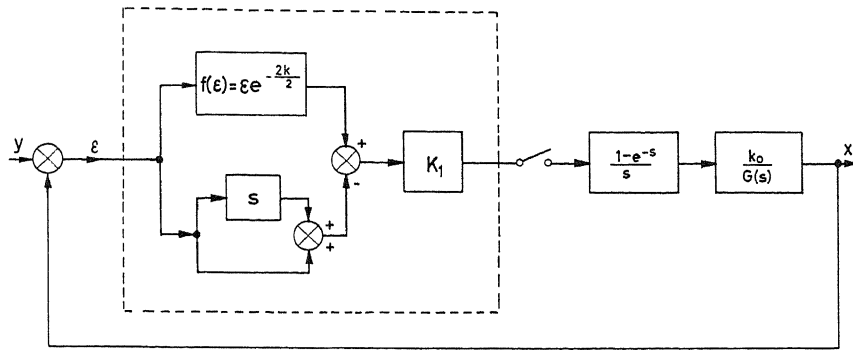


Figure 7. Example of dominant operator adaptive control system

### Example

As an example an adaptive control system has been studied, for which the desired system characteristic  $\varepsilon_1(t + \tau)$  has been simplified and depends on only one variable  $\varepsilon(t)$ .

Let the performance index  $K$  have the form

$$K = \int_0^\infty [\varepsilon(t)]^2 dt$$

Assume also that there is an additional constraint  $K_1$ ,

$$K_1 = \int_{t_0}^\infty (\dot{\varepsilon})^2 dt$$

Using Euler-Poisson equation and Lagrange's method of multipliers, gives

$$^{(2)} - \lambda \varepsilon(t) + \varepsilon(t) = 0$$

Solution of this equation is

$$\varepsilon_1(t) = A e^{-\frac{t}{\sqrt{\lambda}}} + B e^{-\frac{t}{\sqrt{\lambda}}}$$

Assume that

$$\varepsilon_1(\infty) = \varepsilon_1(T) = 0$$

where  $T$  is the time of transient state.

Obtain,

$$\varepsilon_1(t) = \varepsilon(t_0) e^{-\frac{t}{\sqrt{\lambda}}}$$

and constraint equation gives

$$\varepsilon_1(t + \tau) = \varepsilon(t) e^{-\frac{2 K_1 \tau}{[\varepsilon(t)]^2}}$$

Using (12) and (16) we obtain the adaptive control system with dominant operators as shown in Figure 7.

Operation of this system is illustrated in Figures 8 and 9. In Figure 8 are shown error signal responses of this system for different transfer functions of process: Figure 8(a), transfer function of process  $K_0/G(s) = K_0/2s + 1$ ; Figure 8(b), transfer function of process  $K_0/G(s) = K_0/6s + 1$ ; Figure 8(c), transfer function of process  $K_0/G(s) = K_0/s(2s + 1)$ .

In Figure 9 are shown error signal responses of this system for different gain coefficient  $k$  and one transfer function of process  $k_0/G(s) = k_0/s(2s + 1)$ .

Figure 9(a):  $k_1 = 3 k_{opt}$ ; Figure 9(b):  $k_1 = k_{opt}$ ;

Figure 9(c):  $k_1 = 1.5 k_{opt}$

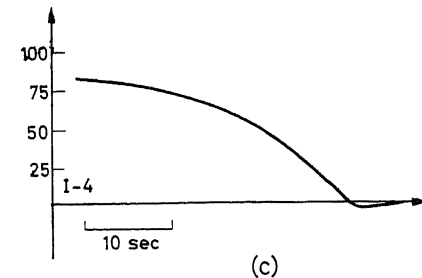
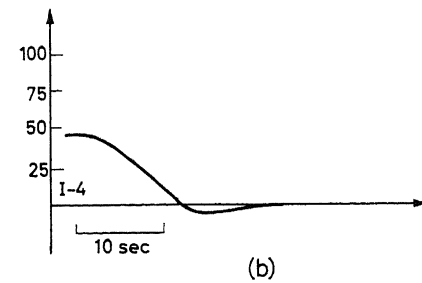
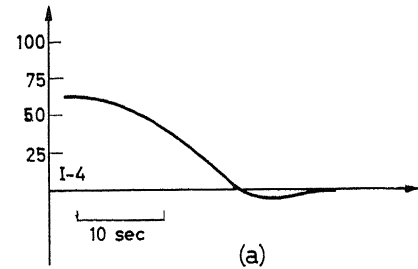


Figure 8. Error signal responses for different transfer function of process:

(a)  $K_0/G(s) = K_0/2s + 1$ ; (b)  $K_0/G(s) = K_0/6s + 1$ ;

(c)  $K_0/G(s) = K_0/s(2s + 1)$

### Conclusions

Dominant operator approach to the synthesis of the adaptive control system has been presented. Time-domain synthesis of a dominant operator has been used. Examples of a dominant operator adaptive control system have been given. This technique supplies a simple method of synthesis for a rapidly adapting control system.

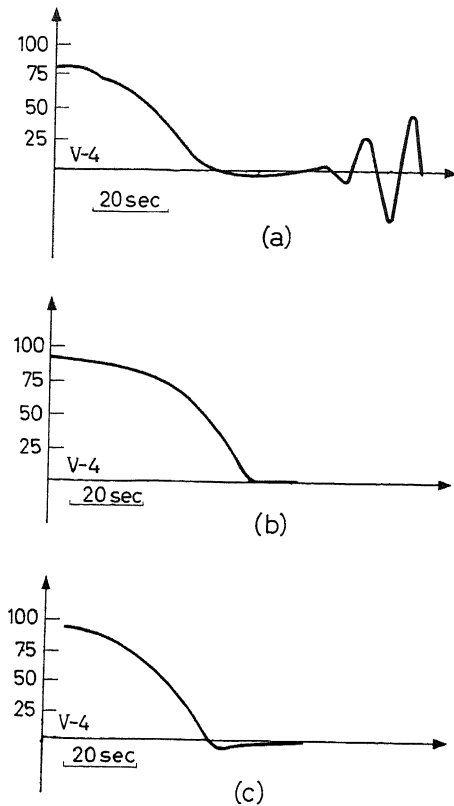


Figure 9. Error signal response for different gain coefficient and one transfer function of process:

- (a)  $k_1 = 3 k_{opt}$ ;  
 (b)  $k_1 = k_{opt}$ ; (c)  $k_1 = 1.5 k_{opt}$

## References

- BELLMAN, R. *Dynamic Programming*. Princeton, 1956
- CHU, Y. Synthesis of feedback control system by phase-angle loci. *Trans. Amer. Inst. elect. Engrs.* Pt II (1952), 330
- KULIKOWSKI, R. Theory of optimal control system. *Arch. Automat. i Telemekh.* (1961), 2-3, (in Polish)
- LI, Y. T., and VANDER VELDE, W. E. Philosophy of non-linear adaptive system. *Automatic and Remote Control*. 577. London; Butterworths
- LI, Y. T., and WITAKER, H. P. Adaptive control systems for transient and frequency response performance. *IFAC Symp. Self-Adaptive System Theory*. Rome, April 1962
- MELLEN, D. L. Application of adaptive flight control. *IFAC Symp. Self-Adaptive System Theory*. Rome, April 1962
- NOLAND, I. H. Stability analysis of rapidly adapting control system. *Amer. Inst. elect. Engrs.* Pap. No. 62-97
- POSPELOW, I. G. Optimal solution in controlling discrete systems in each interval. *IFAC Symp. Self-Adaptive System Theory*. Rome, April 1962
- STAFFIN, R. Executive-controlled adaptive systems. *Applications and Industry*. No. 46 (1961), 523
- STRASZAK, A. Some questions relating to theory and synthesis of self-optimizing automatic control system. *Automatic and Remote Control*. 564. London; Butterworths
- STRASZAK, A. On synthesis some optimal automatic control system (in Russian). *Naucznyje doklady Wyzszej Szkoły Elektromech. i Automat.* No. 4 (1958), 13
- STRASZAK, A. Theory of adaptive control system. *Arch. Automat. i Telemekh.* Nos. 2-3 (1961), 297 (in Polish)
- SCHUCK, O. H. Adaptive flight control. *Automatic and Remote Control*. 645. London; Butterworths
- SCHULTZ, W. C., and RIDEOUT, V. C. Control system performance measures: past, present and future. *IRE Trans. Instn. Radio Engrs. Automat. Contr.* (February 1961), 22

## DISCUSSION

A. R. M. NOTON, *Electrical Engineering Department, University of Nottingham, Nottingham, England*

Unless the additional loop of eqn (17) is added, this is surely not an adaptive system; there appears to be no adjustment of parameters to compensate for either changing process characteristics or changing input statistics.

Furthermore, even if an adaptive loop is included to solve eqn (17) ( $k = h(\tau)/\tau$ ) there is no guarantee of stability. Stability cannot be guaranteed by using only the information  $h_u(\tau)$ , i.e. the value of the impulse response time  $\tau$  ahead.

A. STRASZAK, *in reply*

The property of a control system to be adaptive depends on the behaviour of the system and not on how many additional loops the system has.

Self-adjusting systems may be adaptive or not, because the adjustment is in a certain sense only the technical realization of some non-linear equation. I understand by 'adaptive control system' a system whose behaviour comes close, in an asymptotical way, to the optimal behaviour.

The stability problem for this kind of system must be defined in a different way than for conventional control systems, because with these systems we are interested in its working performance in a finite interval of time, which indeed is usually short.

M. HAMZA, *E.T.H. Zurich, Switzerland*

I read the paper of Dr. Straszak with great interest. In the paper two systems are considered, that of Figure 5 and of Figure 6. To obtain Figure 5 many approximations were made. I would like Dr. Straszak to comment on the performance of this system as compared with the types of adaptive control systems referred to at the beginning of his paper. Further, I feel that the system of Figure 6 will have a long system response time or its performance will not be better than that of the methods referred to in the literature (model and performance index).

A. STRASZAK, *in reply*

In my paper two different adaptive control systems are presented, but only the first of them (Figure 5) is a very fast adaptive control system. To obtain fast adaptation, it is necessary to use the dominant operator approach or a similar technique, because in most approaches we lose much time for measuring the index performance or for comparison with the linear model (usually transfer function).

Published work shows that the additional adaptive loop must be about 1,000 times slower than the main control loop in conventional adaptive control systems.

A. M. HOPKIN, *University of California, Berkeley 4, California, U.S.A.*

Not everyone agrees as to the precise meaning of the term 'adaptive'. Professor Lotfi Zadeh has defined an adaptive control system as one

designed by a person with an adaptive viewpoint. By the Zadeh definition the system under consideration is undoubtedly adaptive.

However, a person accustomed to non-linear controllers sees here a fixed, non-linear control system with sample-and-hold, using feedback to minimize the overall system sensitivity to variations of process dynamics. Originally, linear feedback was introduced for precisely the same purpose. The test of any controller is its degree of success in maintaining the desired output in the face of process variation and external disturbances.

The author defines the desired output in terms of an error criterion with possible constraints, independent of the process dynamics. This implies that the ideal response should be independent of variations of process dynamics also. A measure of the success of the design is the closeness of matching of the actual and ideal responses under various conditions of process variation. *Figure 8* of the paper shows considerable variation of system response with variation of process dynamics. No information is given to permit comparison either with the ideal performance pattern or with behaviour of other controllers that might have been used.

(1) Does the author have comparison data to show that under similar environment this system response is much better than a simpler system, or that this system behaves almost as well as a more elaborate system?

It has been my experience that non-linear control systems have different forms of response for different initial states. In some cases the response will be quite satisfactory for one set of initial error states, but not satisfactory for another set.

(2) How does the example system behave for various initial states, for example various initial error amplitudes?

A. STRASZAK, *in reply*

It is well known that there exist almost a hundred different definitions of the term 'adaptive'; therefore I will not attempt to discuss this problem again.

In the very simple example which I presented for illustration only, the maximum difference between the ideal control system and the dominant operator adaptive control system was less than 5 per cent for different controlled plants.

The dominant operator approach to the synthesis of adaptive systems is based on the state-space technique, therefore this approach is good for all initial states, provided the dimension of the control vector is equal to the dimension of the state vector. This does not follow from the paper as I presented it to this Congress, but from the extension which I have mentioned at this meeting. When we use so-called pseudo-states, then the problem concerning the initial state has to be studied very carefully.

# General Stability Analysis of Sinusoidal Perturbation Extrema Searching Adaptive Systems

V. W. EVELEIGH

## Summary

The stability and response characteristics of extrema seeking adaptive systems, in which sinusoidal perturbations of the adaptive parameter (parameters) provide the derivative action necessary to determine direction and magnitude of the criterion surface gradient at the operating point, are analysed. It is assumed that system variations with time occur slowly as compared with the adaptive loop response time, and that the measured criterion output may be characterized by a general function of the adaptive parameters (usually non-linear) followed by a linear transfer function representing measurement delay. The resulting adaptive control loop is analysed, using modified band-pass to low pass transformation techniques, yielding an equivalent low pass loop in which no perturbation signals appear. This loop is generally non-linear but standard techniques, such as the describing function, may be applied to predict the critical loop gain, the frequency and amplitude of the resulting oscillations, and the required equalization.

Several examples are presented which have been simulated on a Pace analogue computer. Predicted and actual results generally compare within 10 per cent, which is within the usual range of describing function accuracy.

## Sommaire

La stabilité et la réponse de tels systèmes sont analysées. Le système utilise des perturbations sinusoïdales du ou des paramètres adaptatifs, fournissant l'action différentielle nécessaire pour déterminer la direction et l'amplitude du gradient de la fonction-critère. On admet que les variations du système sont lentes par rapport au temps de réponse de la boucle adaptative de commande et que la fonction-critère mesurée à la sortie, peut être représentée par une fonction générale des paramètres adaptatifs (généralement non-linéaire), suivie d'une fonction de transfert linéaire, due au délai de mesure. La boucle adaptative de commande d'un tel système est analysée au moyen d'une théorie modifiée de la transformation passe-bande/passe-bas.

Cette analyse conduit à une boucle passe-bas dans laquelle le signal de perturbation n'apparaît pas.

Cette boucle est en général non-linéaire mais des techniques normalisées telles que la fonction descriptive peuvent être utilisées afin de prédire le gain critique, la fréquence et l'amplitude des oscillations résultantes ainsi que la compensation nécessaire.

Plusieurs exemples, ayant été simulés sur une machine analogique PACE, sont présentés et leurs résultats comparés avec ceux obtenus par la théorie précitée. L'erreur reste dans les 10%, domaine normal de précision d'une fonction descriptive.

## Zusammenfassung

Die Arbeit untersucht die Stabilität und die Übergangsfunktion von Systemen, die selbsttätig nach einem Extremwert suchen. Die sinusförmige Störung der selbstinstellenden Größen ruft eine differentielle Wirkung hervor, die notwendig ist, um Richtung und Größe des Gradienten der (Kriteriums-)Kennfläche im Arbeitspunkt festzulegen. Dabei wird einmal angenommen, daß sich das System, verglichen mit der Übergangszeit des selbstinstellenden Kreises, nur langsam ändert, und zum andern, daß die gemessene Ausgangsgröße durch

eine (meistens nichtlineare) allgemeine Funktion der selbstinstellenden Größen und einer nachfolgenden linearen Übertragungsfunktion, welche der gemessenen Verzögerung entspricht, darstellbar ist. Der sich ergebende selbstinstellende Rückführzweig wird untersucht; durch die Abwandlung eines Bandpasses in einen Tiefpaß ergibt sich ein äquivalenter Tiefpaßzweig, in dem keine Störsignale auftreten. Dieser Zweig ist im allgemeinen nichtlinear, jedoch lassen sich die bekannten Methoden, wie z. B. die Beschreibungsfunktion, zur Vorherbestimmung der kritischen Kreisverstärkung, der Frequenz sowie der Größe der auftretenden Schwingungen und der erforderlichen Kompensation anwenden.

Der Aufsatz enthält einige auf einem PACE-Analogrechner nachgebildete Beispiele. Die Ergebnisse der analytischen und maschinellen Berechnung liegen im allgemeinen innerhalb der Genauigkeit von 10%, was dem Genauigkeitsbereich der Beschreibungsfunktion entspricht.

## Introduction

Although adaptive systems have received much attention during the past decade, analysis of their stability and response characteristics has often been ignored. There are, however, several specific stability analyses available. Morosanov<sup>1</sup> uses the describing function to analyse relay operated peak sensing adaptive controllers; Margolis and Leondes<sup>2</sup> consider the stability of a class of model reference adaptive systems; Schiewe<sup>3</sup> investigates the stability of a three-dimensional adaptive system, assuming the three adaptive loops to be independent and Gibson and Meditch<sup>4</sup> discuss the stability of signal synthesis predictive controllers. The stability and response characteristics of sinusoidal perturbation adaptive systems (hereafter abbreviated SPAS) have been investigated by the author<sup>5, 14, 15</sup>, and the analytical part of that work is presented in this paper.

The study of SPAS is perhaps as old as adaptive control itself. Draper and Li<sup>6</sup>, in a pioneer work, consider an application in which spark timing and fuel mixture were controlled to maximize manifold vacuum, a readily measured even function index of performance (*IP*). The general concept of an even function *IP*<sup>16</sup> is used here to denote any general performance criterion, such as R.M.S. error, which exhibits an extreme value for a specific parameter combination. McGrath and Rideout<sup>7</sup> discuss a system in which two parameters are simultaneously adapted to minimize mean square error. McGrath, Rajaraman and Rideout<sup>8</sup> discuss modifications of this approach, using a model to prevent performance degradation, or hunting loss<sup>9</sup>, due to parameter perturbations. These references provide an excellent background for the subject presented in this paper.

Since all analytical studies depend upon a mathematical system description, it is the purpose of this paper to develop a mathematical model of the general adaptive system and proceed



to simplify that model until an equivalent block diagram is obtained. The block diagram is then used to determine adaptive loop stability and transient response characteristics.

### The Sinusoidal Perturbation Adaptive System (SPAS)

In the interest of brevity only minimum seeking systems are considered here, although the SPAS approach applies equally well to maximum or minimum seeking systems. Assume a one-dimensional SPAS as shown in *Figure 1*. The first step in the analysis of this system is to establish a circuit model for the transfer between parameter input and measured *IP* output, denoted by  $H(s, x, t)$ . In general, steady-state *IP* variations for parameter settings near the optimum value are characterized by a smooth curve, which may often be assumed parabolic<sup>1, 5, 9-13</sup>. Variations in measured *IP* value are not immediately apparent when a parameter adjustment is made, but approach the new value with delay depending upon the measurement process.

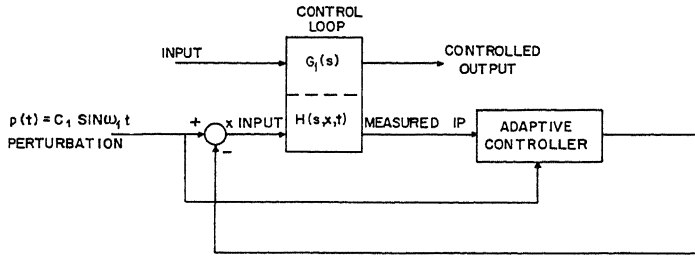


Figure 1. The relationship between the parameter adaptive loop and the standard controller with closed loop transfer  $G_1(s)$

Assume that the *IP* measurement process may be represented by a general function of a parameter  $x$ ,  $f(x)$ , in series with a linear filter representing measurement lag<sup>5, 7</sup>.  $f(x)$  is assumed to possess all derivatives and must provide a well-defined minimum point if the system is to operate properly. Assume that  $f(x)$  is monotonically increasing as  $x$  proceeds from the optimum value in either direction. Thus, if  $x_0$  is the optimum parameter value (yields minimum *IP*),

$$\begin{aligned} f'(x) &> 0, & x > x_0 \\ f'(x) &< 0, & x < x_0 \end{aligned} \quad (1)$$

where  $f'(x)$  denotes the first derivative of  $f(x)$  with respect to  $x$ . An  $f(x)$  curve meeting these requirements is shown in *Figure 2*. In practice, of course, the minimum *IP* and optimum parameter values are not fixed, varying with environment, but this does

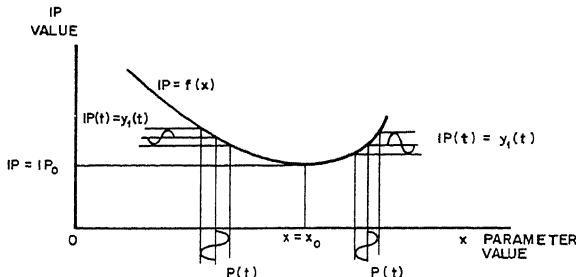


Figure 2. A representative steady state *IP* (performance criterion) variation about the desired operating point,  $x = x_0$ , illustrating the effect of the perturbation signal for + and - parameter errors

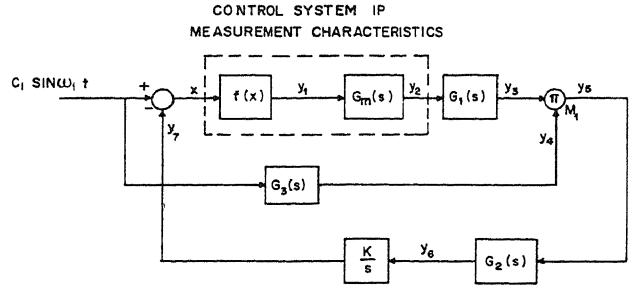


Figure 3. Block diagram representation of a single-dimensional adaptive controller

not appreciably modify the following analysis as long as their time rates of change are small compared to the adaptive response rate, which must be true if the adaptive loop is adequate to compensate for environmental changes.

*Figure 3* is a block diagram of the adaptive control loop. Similar adaptive controllers are discussed in detail in the literature<sup>5-7, 14</sup>, so only a brief steady-state analysis of its operation is presented.  $p(t)$ , as illustrated in *Figure 2*, produces  $y_1(t)$  with an  $\omega_1$  component either in phase with, or  $180^\circ$  out of phase with,  $p(t)$ , depending upon the sign of the parameter error.  $y_3(t)$  is an attenuated and phase shifted version of  $y_1(t)$ .  $G_3(s)$  is chosen so as to give the same steady state-phase shift at  $\omega_1$  as the combination of  $G_m$  and  $G_1$ . Multiplier  $M_1$  is thus a correlation detector, and its average output is either + or - depending upon whether  $y_3(t)$  and  $y_4(t)$  are in phase or  $180^\circ$  out of phase. This, in turn, depends upon whether the offset in  $x$  from  $x_0$  is + or -, respectively.  $G_2(s)$  and the integrator smooth  $y_5(t)$  and develop a correction signal to reduce the error in  $x$ .  $G_1(s)$  and  $G_2(s)$  are chosen to make the loop best conform to design requirements<sup>5</sup>.

Consider an open-loop analysis assuming the loop is broken at the integrator input. Let

$$x = x_1 + C_1 \sin \omega_1 t \quad (2)$$

where  $x_1$  is an arbitrary initial parameter value and  $C_1 \sin \omega_1 t$  is the perturbation signal necessary to make the system operate using an even *IP* function. Using a Taylor Series approximation for the  $f(x)$  curve about  $x = x_1$ ,  $y_1(t)$  becomes

$$y_1(t) = f(x_1) + f'(x_1) \left[ C_1 \sin \omega_1 t \right] + \frac{f''(x_1)}{2!} \left[ C_1 \sin \omega_1 t \right]^2 + \dots \quad (3)$$

If the  $f(x)$  curve is 'smooth' and  $C_1$  is small,

$$\frac{f''(x_1)}{n!} [C_1 \sin \omega_1 t]^n$$

will be negligible for  $n \geq 3$ . Assume that this is the case. Then only the first three terms of (3) need be considered. The third term in (3) may be reduced to a d.c. and a  $2\omega_1$  component. The loop can only pass signals at or near  $\omega_1$ , due to the characteristics of  $G_1$  and  $M_1$ , so it is only necessary to work with the second ( $\omega_1$ ) term of  $y_1(t)$ <sup>5, 7, 14</sup>. Assuming that  $G_m$  and  $G_1$  are linear,

$$y_{3f}(t) = C_1 f'(x_1) A \sin(\omega_1 t + \theta_1) \quad (4)$$

where  $A$  and  $\theta_1$  are the steady-state attenuation and phase shift, respectively, at  $\omega_1$  due to  $G_m$  and  $G_1$ . The subscript  $f$  denotes that

only the fundamental term has been considered.  $G_3$  is chosen such that

$$y_4(t) = B \sin(\omega_1 t + \theta_1) \quad (5)$$

so

$$y_{5f}(t) = \frac{ABC_1 f'(x_1)}{2} [1 - \cos 2(\omega_1 t + \theta_1)] \quad (6)$$

The second term in brackets in (6) will be highly attenuated by  $G_2$  and the integration which follows, or

$$y_6 \cong \overline{y_5} = \frac{ABC_1 f'(x_1)}{2} \quad (7)$$

where the bar denotes time average.  $ABC_1$  is positive, so  $y_6$  will tend to reduce parameter offset when the loop is closed, due to the limits imposed by (1) on  $f'(x)$ . If  $f(x)$  is parabolic near the optimum, it is easily shown<sup>5, 7, 14</sup> that (7) reduces to

$$y_6 \cong \overline{y_5} = a(x_1 - x_0) \quad (8)$$

where  $a$  is the gain, indicating that loop operation is quasi-linear for a parabolic  $IP$  function. In practice, it should always be possible to approximate the  $IP$  variation as accurately as desired in some  $\varepsilon$  neighbourhood of the optimum by a parabolic curve (surface in the  $n$ -dimensional case). Thus, if one is interested only in defining the small scale asymptotic stability and response conditions of the loop, the analysis is considerably simplified.

The previous analysis was carried out assuming the loop to be open so that steady-state sinusoidal techniques could be applied. When the loop is closed this analysis is no longer strictly valid. If, however, loop response is slow, relative to the time required to establish steady-state conditions, as it normally is in practice<sup>5, 7</sup>, this analysis closely approximates the actual performance. In any event, it provides a heuristic argument to explain why the loop operates satisfactorily.

### Derivation of an Equivalent Low Pass Adaptive Loop

Modified describing function and band-pass to low-pass transformation techniques may be applied to the adaptive loop of Figure 3 to derive an equivalent low pass network. This is the key step in stability study of SPAS. The perturbation signal is treated as a carrier, and equivalent low pass networks are derived for each of the linear and non-linear loop elements between perturbation input (modulation) and correlation detector output (demodulation).

Let

$$x(t) = x_0 + C_1 \sin \omega_1 t + C_m \sin(\omega_m t + \theta_m) \quad (9)$$

where  $C_m$ ,  $\omega_m$  and  $\theta_m$  are the amplitude, frequency and phase, respectively, of a circulating natural frequency component. The  $\omega_m$  term is the sinusoidal excitation term normally assumed in describing function analysis. Assume that  $\omega_m \ll \omega_1$ , which is true if adequate smoothing ( $G_1$  and  $G_2$ ) is used to assure reasonable loop transient response<sup>5, 14</sup>. (The analysis, as illustrated in Tables 1 and 2, yields accurate results even as  $\omega_m \rightarrow \omega_1$ , which represents a severe test of the system's equivalent network derived herein.) The  $\omega_1$  component of  $y_1(t)$  is obtained from (2) and (3), assuming that  $C_m \sin(\omega_m t + \theta_m)$  changes but little during one cycle of  $\sin \omega_1 t$ , as

$$y_{1f}(t) = C_1 f'[x_0 + C_m \sin(\omega_m t + \theta_m)] \sin \omega_1 t \quad (10)$$

$C_1 f'[x_0 + C_m \sin(\omega_m t + \theta_m)]$  product modulates the  $\sin \omega_1 t$  (carrier) term. The modulation envelope has a frequency component at  $\omega_m$ , and its amplitude may be determined by finding the Fourier coefficient of the output  $\omega_m$  term which results if  $x_0 + C_m \sin(\omega_m t + \theta_m)$  is applied to  $f'(x)$ . Evaluation of  $A(C_m)$ , the describing function for  $f'(x)$ , is thus the first step in development of the equivalent low pass loop. Then

$$y_{1f}(t) = A(C_m) \sin \omega_1 t \sin(\omega_m t + \theta_m) \quad (11)$$

or

$$y_{1f}(t) = \frac{A(C_m) C_1}{2} \{ \cos[(\omega_1 - \omega_m)t - \theta_m] - \cos[(\omega_1 + \omega_m)t + \theta_m] \} \quad (12)$$

Bandpass to low pass transformation techniques may now be applied to find equivalent low pass filters for  $G_1$  and  $G_m$ . As an example of this process, assume that

$$G_m(s) = \frac{K_m}{\tau_m s + 1} \quad (13)$$

and

$$G_1(s) = \frac{\omega_1^2 s}{s^2 + \omega_1 s + \omega_1^2} \quad (14)$$

These are reasonable choices in the light of experimental evidence<sup>5, 14</sup>, but the approach is by no means limited to these filters.  $G_3$  may now be chosen as

$$G_3(s) = \frac{K_3}{\tau_m s + 1} \quad (15)$$

since  $G_1$  gives no phase shift at  $\omega_1$ . Any  $G_3$  giving the same phase shift could be used.

Let the phase shift of  $G_m(j\omega_1)$  be  $\theta_1$ , and assume that the phase characteristic near  $\omega_1$  is sufficiently linear that the phase shift at  $s = j(\omega_1 - \omega_m)$  may be represented by  $\theta_1 + \Delta\theta$  and that at  $s = j(\omega_1 + \omega_m)$  by  $\theta_1 - \Delta\theta$ . Let the gain of  $G_m(j\omega_1) = a_1$ . Thus

$$a_1 = \frac{K_m}{[1 + (\tau_m \omega_1)^2]^{\frac{1}{2}}} \quad (16)$$

Assume further that the attenuation at  $s = j(\omega_1 - \omega_m)$  and  $s = j(\omega_1 + \omega_m)$  is approximately  $a_1$ . This is reasonable for  $\omega_m \ll \omega_1$ . The steady-state output of  $G_m(s)$  thus becomes

$$y_2(t) = \frac{a_1 A(C_m) C_1}{2} \{ \cos[(\omega_1 - \omega_m)t - \theta_m + \theta_1 + \Delta\theta] - \cos[(\omega_1 + \omega_m)t + \theta_m + \theta_1 - \Delta\theta] \} \quad (17)$$

which may be rewritten as

$$y_2(t) = a_1 A(C_m) C_1 [\sin(\omega_1 t + \theta_1) \sin(\omega_m t + \theta_m - \Delta\theta)] \quad (18)$$

$G_m$  thus shifts the carrier  $\theta_1$  rad and the intelligence signal  $-\Delta\theta$  rad. Only the intelligence signal shift is significant in the equivalent low pass loop.

An equivalent transfer function may now be defined such as to approximate the phase characteristics of the original network, or

$$G_{me} = \frac{a_1}{\alpha s + 1} \quad (19)$$

$\alpha$  is chosen such that

$$\left. \frac{d\theta}{d\omega} \right|_{\omega=0} = -\alpha = \left. \frac{d\theta}{d\omega} \right|_{\omega=\omega_1} \quad (20)$$

Thus the phase slope of the low pass equivalent network about  $\omega = 0$  must approximate that of the carrier network about  $\omega = \omega_1$ . It should be noted that the approach used here is somewhat more general than the usual bandpass to low pass transformation, applying to all networks whether band centred at  $\omega_1$  or not. Of course the accuracy is best when dealing with high  $Q$  bandpass networks centred at  $\omega_1$ .

Derivation of  $G_{1e}(s)$  is carried out using similar techniques, so need not be discussed in detail. The approach applies even if  $G_1(s)$  is not band centred, as illustrated later with an example. The equivalent loop may be revised, if desired, to obtain greater accuracy by using a value of  $\Delta\omega$  close to the natural frequency of the system to determine the value of  $\alpha$  in the equivalent low pass network, i.e., instead of evaluating  $\alpha$  from (20), use

$$-\alpha = \left. \frac{\Delta\theta}{\Delta\omega} \right|_{\omega=\omega_1}$$

where  $\Delta\omega$  is near the natural frequency of the loop. This has been found to yield a significant increase in accuracy.

The phase reference input to  $M_1$  is given by

$$y_4(t) = B \sin(\omega_1 t + \theta_1) \quad (21)$$

and the signal input is

$$y_{3f}(t) = \mu \sin(\omega_1 t + \theta_1) \sin(\omega_m t + \theta_m - \gamma) \quad (22)$$

where  $\mu$  is the amplitude,  $\theta_1$  and  $\gamma$  the phase lags, which appear after passing through  $G_m$  and  $G_1$ . Multiplying (21) by (22) gives the  $M_1$  output as

$$y_{5f}(t) = \mu B \sin^2(\omega_1 t + \theta_1) \sin(\omega_m t + \theta_m - \gamma) \quad (23)$$

which may be rewritten as

$$y_{5f}(t) = \frac{B\mu}{2} \sin(\omega_m t + \theta_m - \gamma) [1 - \cos 2(\omega_1 t + \theta_1)] \quad (24)$$

The first term obtained by expanding (24) is the desired signal and the second is a relatively insignificant high frequency component. The describing function for  $M_1$  is the ratio of signal out to envelope amplitude in, or

$$G_{m1} = \frac{B}{2} \quad (25)$$

The remaining elements in the loop act upon the signal terms directly, so no further transformations are required.

#### Example 1

Let

$$C_1 = \omega_1 = 10, \tau_m = 0.1, B = 0.707 \text{ and } K_m = 1$$

Assume that

$$f(x) = IP = k_1(x - x_0)^2 + k_2 \quad (26)$$

adequately approximates the steady-state  $IP$  curve near the optimum. Then

$$f'(x) = 2k_1(x - x_0) \quad (27)$$

and (10) becomes

$$y_{1f}(t) = 2k_1 C_m C_1 \sin \omega_1 t \sin(\omega_m t + \theta_m) \quad (28)$$

$A(C_m)$  is thus  $2k_1 C_m$  and the 'gain' associated with  $f(x)$  is  $2C_1 k_1 = 20k_1$ . Letting  $k_1 = 0.01$ ,  $20k_1 = 0.2$ .  $\alpha$  is found from (20) to be 0.05 and  $a_1 = 0.707$ , so

$$G_{me}(s) = \frac{0.707}{0.05s + 1} \quad (29)$$

Similar calculations show that

$$G_{1e}(s) = \frac{10}{0.2s + 1} \quad (30)$$

Combining these equivalent low pass networks yields a loop transfer

$$G_{Le}(s) = \frac{0.5 K G_2(s)}{s(0.05s + 1)(0.2s + 1)} \quad (31)$$

This  $G_{Le}(s)$  has been used to predict critical gain levels  $K_u$  and oscillation frequencies  $\omega_0$  for various choices of  $G_2(s)$ . The results are compared, in Table 1, to simulation results obtained

Table 1. Predicted versus actual instability conditions.  $K_u$  is the critical value of the  $K$  defined in Figure 3.  $G_1(s)$  is as defined by eqn (14) with  $\omega_1 = 10$  in each case

$G_2(s)$	$\omega_1$	Predicted		Actual	
		$\omega_0$	$K_u$	$\omega_0$	$K_u$
$\frac{1}{(0.2s + 1)^2}$	10	2.8	8.9	3	8.5
$\frac{1}{(0.5s + 1)^2}$	10	1.45	4.6	1.47	4.8
$\frac{1}{(0.2s + 1)^2}$	20	3.5	22	3.92	25
$\frac{1}{(0.5s + 1)^2}$	20	1.7	12	1.97	14

using an analogue computer. Results presented for  $\omega_1 = 20$  were obtained by deriving new equivalent transfers for  $G_m(s)$  and  $G_1(s)$ , but  $G_1(s)$  was left centred at  $\omega_1 = 10$ . This represents a severe test of the general approach.

#### Example 2

Consider an  $IP$  variation defined as follows:

$$\left. \begin{aligned} f(x) &= \frac{1}{50} x^2 & -25 \leq x \leq 25 \\ f(x) &= |x| - 12.5 & |x| > 25 \end{aligned} \right\} \quad (32)$$

For this  $IP$  function the carrier signal amplitude increases linearly with parameter error until the error reaches  $\pm 25$  and remains constant as error increases beyond that point. The equivalent loop for  $C_1 = \omega_1 = 10$  and  $\tau_m = 0.1$  is shown in Figure 4.  $f'(x)$  is the standard saturation function, as readily derived from (32).

Predicted and simulated values of critical gain and natural frequency for this system are compared in Table 2. Figure 5 shows curves comparing predicted and actual limit cycle amplitudes versus  $K$  for three choices of  $G_2(s)$ .

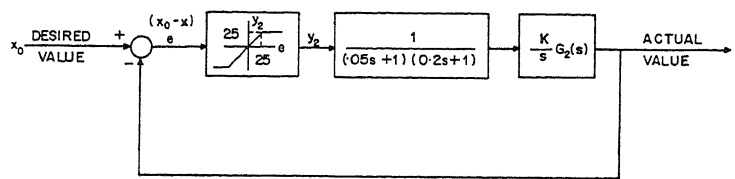


Figure 4. The equivalent low pass loop for Example 2

Table 2. Predicted versus actual instability conditions for the network shown in Figure 4

$G_2(s)$	$\omega_1$	Predicted		Actual	
		$K_u$	$\omega_0$	$K_u$	$\omega_0$
$\frac{1}{(0.1s + 1)^2}$	10	6	4	6.4	3.7
$\frac{1}{(0.5s + 1)^2}$	10	2.3	1.45	2.4	1.43
$\frac{1}{(s + 1)^2}$	10	1.33	0.8	1.45	0.83
$\frac{1}{(2s + 1)^2}$	10	0.83	0.45	0.88	0.44

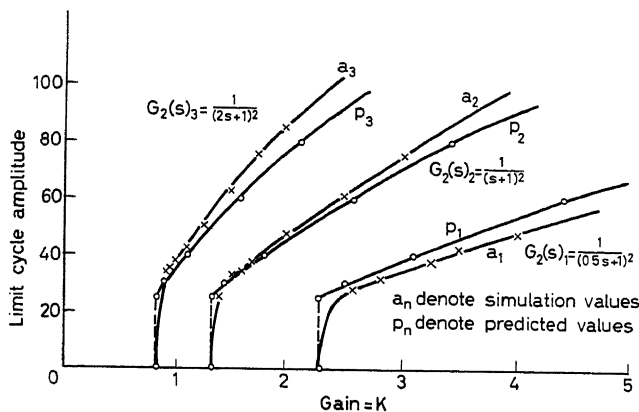


Figure 5. Comparison of predicted and actual limit cycle amplitudes for Example 2

### Conclusions

A general method for determining the stability and response characteristics of single-dimensional SPAS has been developed. This method depends only upon the availability of a mathematical or graphical approximation of the *IP* function used to develop the adaptive error signals. A method for deriving an equivalent low pass control loop has been outlined and examples presented to illustrate its use. If the *IP* is parabolic in  $x$ , the equivalent loop is linear, but a series non-linear element is required to represent more complex *IP* functions. The method has been tested using analogue computer simulations which show predicted and simulated results to agree within 10 per cent in most cases. The method thereby yields results within the accuracy normally anticipated from non-linear analysis techniques such as the describing function.

Derivation of the equivalent low pass has been emphasized in this paper. Additional analyses use standard techniques, so have not been discussed in detail. The method is clear. Determine or approximate  $f'(x)$  in the region of interest. (A straight line approximation should usually be adequate.) Find the describing function for this  $f'(x)$ . Determine the equivalent low pass networks for  $G_m$  and  $G_1$ . Determine the detector gain. Analyse the resulting control loop.

In those multi-dimensional cases where there is little intercoupling between the adaptive loops, this approach may be applied to each of the loops in turn to predict their response characteristics and instability conditions. Allowance of somewhat more than the usual gain and phase margins in the individual loops generally yields satisfactory overall multi-dimensional adaptive response. Loop intercoupling can be minimized by using narrow band filters for the  $n$   $G_1(s)$  units required. Narrow band filters give more phase lag and thus yield slower response, so a compromise is required. Intercoupling may also be reduced by careful choice of the adaptive parameters. No general method has yet been developed for the consideration of  $n$ -dimensional systems with significant intercoupling between the adaptive loops. This problem deserves further attention.

Although no justification is attempted, the equivalent low pass loop derived herein may be used as a guide for design equalization of SPAS loops to give a desired damping factor, peak overshoot, bandwidth, etc. Further work is being done in these areas and many interesting problems have yet to be considered.

### References

- MOROSANOV, I. S. Methods of extremum control. *Automat. Remote Contr.* 18 (1957), 1077-1092
- MARGOLIS, M., and LEONDES, C. T. A parameter tracking servo for adaptive control systems. *Trans. Inst. Radio Engrs Automat. Contr.* PGAC (Nov. 1959)
- SCHIEWE, A. J. *A Three Dimensional Adaptive Control System*. Ph. D. Thesis, School E. E., Purdue U., W. Lafayette, Ind., (June, 1960)
- GIBSON, J. E., and MEDITCH, J. S. *A Class of Predictive Adaptive Controls*, PRF 2358, Proj. 8225, Task 82181, School of E. E., Purdue U., W. Lafayette, Ind. (Feb. 1961)
- EVELEIGH, V. W. *A Comparison of Two Approaches to Extrema Searching Adaptive Systems*. Ph. D. Thesis, School of E. E., Purdue U., W. Lafayette, Ind. (June, 1961)
- DRAPER, C. S., and LI, Y. T. Principles of optimizing control systems and an application to the internal combustion engine. *Amer. Soc. mech. Engrs, N.Y.* (1951)
- MCGRATH, R. J., and RIDEOUT, V. C. A simulator study of a two parameter adaptive system. *Trans. Inst. Radio Engrs Automat. Contr.* PGAC (Feb. 1961)
- MCGRATH, R. J., RAJARAMAN, V., and RIDEOUT, V. C. A parameter perturbation adaptive control system. *Trans. Inst. Radio Engrs Automat. Contr.* PGAC (May, 1961)

- <sup>9</sup> TSIEN, H. S. *Engineering Cybernetics*. 1954. Ch. 15, p. 217 New York; McGraw-Hill
- <sup>10</sup> FELDBAUM, A. A. Automatic optimizer. *Automat. Remote Contr.* Aug. (1958), 718-728
- <sup>11</sup> FELDBAUM, A. A. *Computational Methods in Automatic Control* (Russian) 1959. Moscow; State Publishing House of Physical and Mathematical Literature
- <sup>12</sup> STAKHOVSKII, R. I. A multichannel automatic optimizer for solving variational problems. *Automat. Remote Contr.* Nov. (1959), 1435-1445
- <sup>13</sup> STAKHOVSKII, R. I. Twin channel automatic optimizer. *Automat. Remote Contr.* Aug. (1958), 729-740
- <sup>14</sup> EVELEIGH, V. W. General stability analysis of sinusoidal perturbation extrema searching adaptive system. *GETIS No. R61ELS-115*, E. Lab. Gen. Elect. Co., Syracuse, N. Y.
- <sup>15</sup> EVELEIGH, V. W. A survey of adaptive control technology. *GETIS No. R61ELS 142*, E. Lab., Gen. Elect. Co., Syracuse, N. Y.
- <sup>16</sup> NEWTON, G. C. Jr., GOULD, L. A., and KAISER, J. F. *Analytical Design of Linear Feedback Controls*. p. 31. 1957. New York; Wiley

## DISCUSSION

O. L. R. JACOBS, *Kings Buildings, Mayfield Road, Edinburgh 9, Scotland*

The author has presented an interesting approach to the analysis of sinusoidal perturbation extremum control systems.

In recent years a variety of extremum control systems have been described<sup>1</sup>, and one outstanding problem is to compare their performances. For this purpose the general formulation of an extremum control system shown in *Figure A* might be adopted.

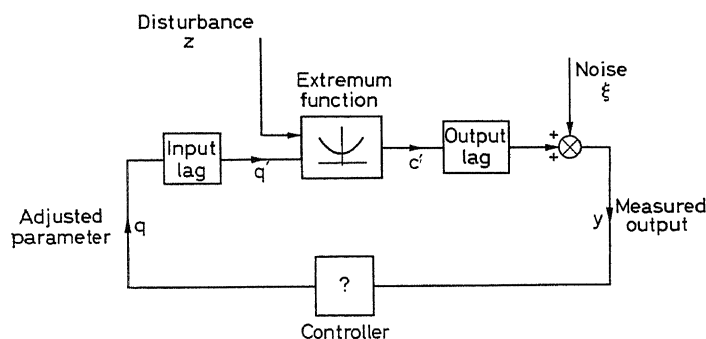


Figure A. Extremum control system ( $q, q', z$  may be vector quantities)

In order to avoid triviality in certain examples it is necessary for the general formulation to include three factors: random disturbances  $z$  which make continuously acting extremum control necessary; and random measurement noise  $\xi$  and dynamic lags which are the two factors limiting performance.

Controllers that have been described in the literature use extremum control strategies that fall into two categories; those using perturbations such as described here, and those using finite steps<sup>2</sup> such as described in the paper by J. L. Douce. Intuitive reasoning suggests that a perturbation strategy should be used where noise is the dominant factor limiting performance, and that a stepping strategy should be used where dynamic lags dominate performance. Unfortunately it has not yet been possible to make any comparison for the two types of strategy controlling the same system.

Published analyses of extremum control systems tend to neglect at least one, and often two, of the important factors (disturbances, noise, lags). Analysis of perturbation systems often neglects noise, analysis of stepping systems often neglects dynamics, analysis of both types often neglects disturbances.

In the present paper both noise and disturbances are neglected, and one output lag is considered. I would like to ask the author whether he thinks it possible to extend his approach to analyse the performance of a perturbation system in which noise and disturbances are included, so that it may be possible to make comparisons with stepping systems.

## References

- <sup>1</sup> JACOBS, O. L. R. Hill climbing. *Control* (Feb. 1962) 92
- <sup>2</sup> JACOBS, O. L. R. and WONHAM, W. M. Extremum control in the presence of noise. *J. Electron. Contr.* Vol. II, No. 3 (1961) 193

V. W. EVELEIGH, *in reply*

I thank Dr. Jacobs for his interesting comments and his question regarding the effect of noise upon loop performance. Noise effects have, in a limited sense, been considered in the analysis, since they determine the smoothing time constants required in the *IP* measurement circuit. This is not the sense referred to by Dr. Jacobs, however.

I feel that the analysis can be extended to include noise and transient disturbances, thereby allowing a fuller comparison with stepping systems. In fact, a comparison between the transient response characteristics of SPAS and discrete systems has been carried out (see Reference 5 of the paper) using a model *IP* surface, measurement lag, and disturbances. Only a limited noise analysis was carried out, however. Our current plans are for further study of this problem, including recent developments in the technology, and I hope to report the results in detail at a later conference.

J. D. ROBERTS, *77, High Street, Cherry Hinton, Cambridge, England*

The stability of an SPAS is very important and it is significant that such realistic results have been obtained with this theory. It must be mentioned, however, that stability investigations need to be different in linear and non-linear systems. A linear system, if it is asymptotically stable in one state, will converge to that state from any state. A non-linear system may, however, have bounded regions of stability in its phase space. If an SPAS is to respond to Gaussian random disturbances, it must be stable for all values of its phase coordinates; but no value of  $K$  in a simple SPAS system of the type described makes the system stable everywhere; as shown in *Figure A*.

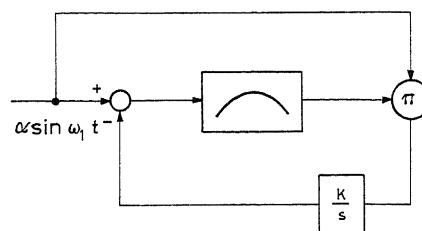


Figure A

The system has two phase coordinates  $\theta, x$  which obey the following equations:

$$\left. \begin{aligned} \frac{d\theta}{dt} &= \omega_1 \\ \frac{dx}{dt} &= -K\alpha \sin \theta (\alpha \sin \theta + x)^2 \end{aligned} \right\}$$

Omitting the most tedious part of the mathematics, we note merely that a differential equation of the form

$$\frac{dx}{dt} = f(t)x^2 \quad \text{where } f(t) > f_0 > 0$$

is Laplace unstable, i.e. it goes to  $\infty$  in a finite time. Moreover, this finite time can be made arbitrarily small by making the initial value of  $x$  sufficiently large. Since  $\sin \theta$  is positive for a finite time, the system is unstable for some phase coordinates for any value of  $K$ .

It would be ideal to have an SPAS which was stable everywhere.

An SPAS of the type shown in Figure B will converge to an unknown hill of known unit curvature from any point of the phase coordinates.

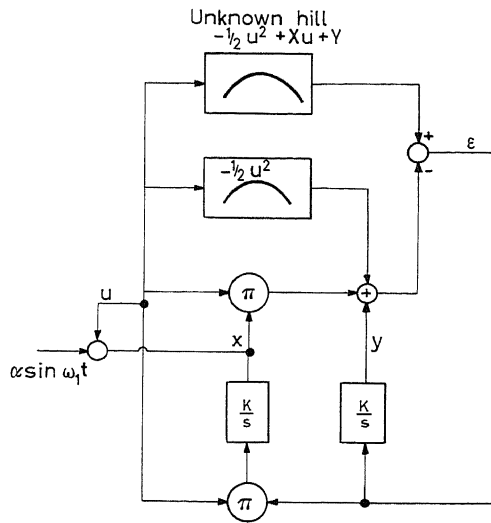


Figure B

We have the equations:

$$u \equiv x + \alpha \sin \omega_1 t$$

$$\epsilon \equiv (X - x)u + Y - y$$

$$\frac{dx}{dt} = K\epsilon u$$

$$\frac{dy}{dt} = K\epsilon$$

Now consider the Liapunov function

$$L = (X - x)^2 + (Y - y)^2$$

which satisfies:

$$\begin{aligned} \frac{dL}{dt} &= -2(X - x)\frac{dx}{dt} - 2(Y - y)\frac{dy}{dt} \\ &= -2K\epsilon[(X - x)u + Y - y] \\ &= -2K\epsilon^2 < 0 \end{aligned}$$

V. W. EVELEIGH, *in reply*

I thank Dr. Roberts for his interesting example illustrating potential instability regions for adaptive systems of the type discussed in this paper. I agree that such systems, depending upon the form of the *IP* surface, are often unstable for sufficiently large initial conditions. This defines a region in the phase space which, it is interesting to note, because of the sinusoidal perturbation signal, has time-varying boundaries outside of which the system is unstable. It was found in some of the experimental work upon which the paper is based that attempts to initiate action from large values of  $x$  often drove the

computer hard into saturation, which is similar to the phenomenon suggested by Dr. Roberts. In all cases considered, however, it was found that a reasonably large region in the parameter space was unconditionally stable for properly chosen gain levels.

J. L. DOUCE, *Queen's University, Belfast, N. Ireland*

I disagree strongly with the analysis presented in this paper, applied to practical optimizing control systems. The author assumes that performance index, such as R.M.S. error, can be related to parameter value by a non-linear characteristic followed by a linear filter. This is not so, except in the case of theoretical interest only, when the error is a constant with no variation with time whatsoever. In any control system the error fluctuates, and to obtain any measure of the change in performance index due to a parameter perturbation we must, from statistical limitations, measure R.M.S. error over a time period very much longer than the response time of the system.

The author's values used in the examples are quite unrealistic. The theory applies more to the analysis of particular non-linear feedback systems than to self-optimizing control systems.

I agree with the author that the response time of this system is slow. In practice, however, increasing the gain produces oscillations at perturbation frequency, not at some unrelated lower frequency. If a clamp or sample and hold is inserted in the optimizing loop, the gain can be increased by a factor of about 30, then the speed of response is correspondingly increased, and parameter fluctuations eliminated. Stability analysis of this new system is relatively simple.

As a minor point, the author suggests that the phase lag of the control system is cancelled by a network  $G_3(j\omega)$ . With high perturbation frequencies this is just not possible, since the phase shift in the control system varies with parameter value. Incorrect phasing causes the system to settle to a false minimum, and can give instability in extreme cases.

V. W. EVELEIGH, *in reply*

I thank Dr. Douce for his list of comments which will help me to answer questions that may have arisen in other people's minds as well. First of all, regarding the representation of *IP* measurement process as a general non-linear function in series with a linear filter, this approximation is consistent with the usual methods used in the literature. It is assumed throughout the development that the *IP* function is a random variable with expected value in steady state as described by the non-linear curve. The filter is included to represent the smoothing required to yield a reasonably steady output approximating the mean value of this random variable, which may be R.M.S. error, manifold vacuum, or some similar signal. It is true that, for the usual *IP*'s and *IP* measurement techniques used, the measurement time constant (or time constants) is significantly longer than the time constants of the control loop itself. Thus adaptive loop response is generally slow relative to that of the control loop. I feel that this representation of a practical *IP* measuring device acting upon manifold vacuum, for example, is quite reasonable. Thus the practical significance of the work would not seem to be appreciably degraded by this assumption.

The effect of noise at the measured *IP* output upon system performance may readily be determined, either analytically or experimentally. If the noise is uncorrelated with the perturbation signal, it has no average effect upon loop response. The parameter value will be disturbed in an R.M.S. sense from its optimum setting by an amount proportional to the noise level. However, this does not alter the general stability characteristics of the loop in most cases of interest and particularly not in either example presented in the paper.

The values used in the examples were chosen primarily for convenience in the computer simulation of the system, and not with a specific application in mind.

It is true that in many cases increasing adaptive loop gain sufficiently will produce oscillations at the perturbation frequency, but, as illustrated by the data from Examples 1 and 2 in the paper,

oscillations at lower frequencies often occur for much lower loop gain values. The analysis predicts accurately the frequency of this oscillation and the gain at which it begins. As pointed out by Dr. Douce, the use of a sample and hold filter offers interesting possibilities for reducing response time and improving loop stability.

I see no reason why the response lag of the  $IP$  measurement process should consistently vary with parameter value, it generally depending primarily upon the time constants chosen for the smoothing filters used to extract an approximation to the mean value of the  $IP$  function. If it does change, however, due to some other phenomenon, I can only suggest that  $G_3(j\omega)$  should be chosen to give the best compromise phase shift.

Finally, the limit cycle exhibited in Example 2 is a direct result of the saturation non-linearity and frequency characteristics which appear in the low pass equivalent loop for that example, shown in Figure 4 of the paper. Standard analysis of that loop using describing function techniques yields limit cycle amplitude as a function of loop gain, as shown in Figure 5.

R. K. SMYTH, 1960 Domingo Road, Fullerton, California, U.S.A.

Dr. Eveleigh presents an interesting method for deriving the equivalent low pass transfer function for the bandpass filter contained in the adaptive adjustment loop.

The discussor has considered a similar problem and has derived a somewhat different approximation which also gives good experimental correlation with the predicted stability. The bandpass filter considered was:

$$\frac{2\zeta\omega_n s}{s^2 + 2\zeta\omega_n s + \omega_n^2}$$

which is similar to the bandpass filter in the paper's eqn (14).

For this transfer function it was shown that a good approximation for the low pass characteristics of the bandpass filter is given by:

$$G_{1e}(s) \cong \frac{\zeta\omega_n}{s + \zeta\omega_n} \cdot \frac{\alpha\zeta\omega_n}{s + \alpha\zeta\omega_n}$$

where  $\alpha \cong 4$  for low values of the damping ratio  $\zeta$ . Note that the first term is the one obtained with conventional a. c. servo theory.

Would the author like to comment on this problem?

#### Reference

SMYTH, R. K. and NAHI, N. E. Phase and amplitude sinusoidal dither adaptive control system. *Proc. Jacc.* (June 1963) Minneapolis, Min. U.S.A.

V. W. EVELEIGH, *in reply*

I thank Dr. Smyth for his suggestion of an alternate representation of the bandpass filter to more accurately account for the lack of arithmetic symmetry about the centre frequency.

Dr. Smyth's suggestion may provide an efficient way to improve accuracy while bypassing the iteration required in the incremental improvement technique presented in the paper. I plan to compare the resulting accuracies of these approaches. I would like to mention that the low pass filter  $G_m(s)$  is likewise asymmetrical about the perturbation (carrier) frequency and gives rise to similar difficulties.

P. EYKHOFF, *Technological University, Electrical Engineering Department, Delft, Netherlands*

Dr. Eveleigh discusses the case of an optimum searching system based on the relation

$$\frac{d(IP)}{dt} = \frac{\partial(IP)}{\partial t} = \frac{\partial(IP)}{\partial x} \frac{dx}{dt}$$

where, in his notation,  $IP$  is the measured index of performance and  $x$  is the parameter (set point) to be adjusted.

If  $dx/dt$  is a known test signal and if  $d(IP)/dt$  can be measured then information on the 'gradient'  $\partial(IP)/\partial x$  can be derived.

The study of these types of systems is a timely subject. This is indicated by the fact that at this conference, besides the paper at hand, there are also two other presentations, by Alimov, and Isobe and Totani, that discuss such topics.

In this discussion I would like to indicate briefly another approach to the analysis problem<sup>1, 2</sup>. This approach has the advantage that it does not require the assumption made in connection with eqn (9), which amounts to a rather slow convergence of the adjusting scheme.

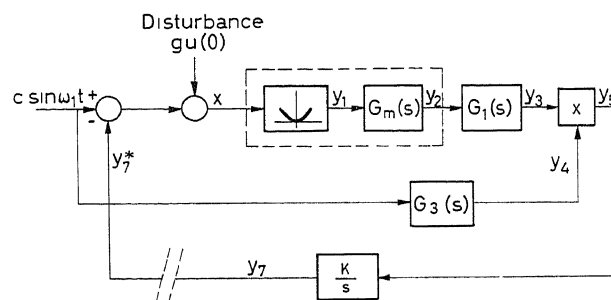


Figure A

Consider Figure A which is almost identical with Figure 3 of the paper except for the disturbance  $gu(0)$  with

$$\begin{aligned} u(0) &= 0 \quad \text{for } t < 0 \\ u(0) &= 1 \quad \text{for } t \geq 0 \end{aligned} \quad (1)$$

which represent the offset from  $t = 0$  that has to be compensated for by the adaptive system. The index of performance has been chosen as a quadratic one.

For the moment assume

$$y_7^* = g(1 - e^{-\alpha t})u(0) \quad (2)$$

then

$$x = C \sin \omega_1 t + g e^{-\alpha t} u(0) \quad (3)$$

and

$$y_1 = \{C \sin \omega_1 t + g e^{-\alpha t} u(0)\}^2 \quad (4)$$

One can take the Laplace transform of eqn (4); Figure B (i) gives the poles of this transformation. The zeros can be found in the known way taking into consideration the amplitude ratio.

Assume the dynamics of  $G_m(s)$  and  $G_1(s)$  to be given by the pole zero pattern of Figure C, then the poles of signal  $y_3$  are found in Figure B (ii). With the aid of the real multiplication theorem

$$L[f_1(t) \cdot f_2(t)] = \frac{1}{2\pi j} \int_{c-j\infty}^{c+j\infty} F_1(s-v) F_2(v) dv \quad (5)$$

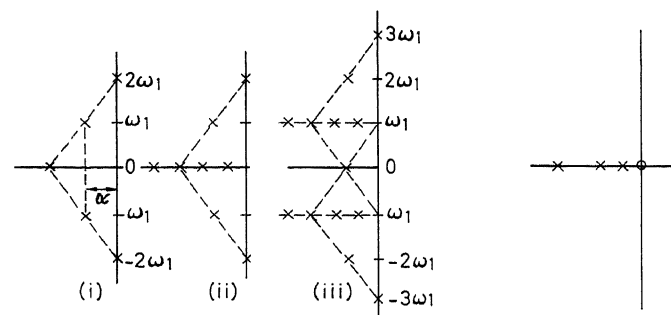


Figure B

Figure C

it can be shown that the poles of the signal  $y_5$  can be found by shifting the pole pattern of Figure B (ii) according to  $j\omega$  and  $-j\omega$  respectively and adding in the  $s$  plane. This leads to Figure B (iii).

The integrator adds a pole in the origin, to give the pole patterns of  $y_7$ . Both poles on the real axis amount to a time function

$$h(1 - e^{-at})u(0) \quad (6)$$

If  $g = h$  then our initial assumption (2) for  $y_7^*$  was right. Now an integrator constant  $K$  can be found for which this is true. The poles off the real axis have been neglected in passing from  $y_7$  to  $y_7^*$ . It is simple to show that the residues of these poles are at least an order of magnitude smaller than the residues for the poles on the axis.

I found it very interesting to compare this treatment of the problem with the method of the low pass equivalent network.

There remains one question. In experimental work I never found a need for an additional low pass filter  $G_2(s)$ ; even for high speeds of convergence the correction signal  $y_7$  was sufficiently smooth. So what is the reason for introducing  $G_2(s)$  and accepting with it the chance of instability?

#### References

- <sup>1</sup> EYKHOFF, P. Optimizing control and process-parameter estimation. *Ph. D. thesis, Elect. Engng.* Univ. of California, Berkeley (Dec. 1960)
- <sup>2</sup> EYKHOFF, P. and SMITH, O. J. M. Optimizing control with process-dynamics identification. *Pap. IRE No. 60 AC-15*, Joint Autom. Control Conf., Cambridge, Mass. (Sept. 1960); *IRE Trans.*, Vol. AC-7, No. 2, (1962) 140-155

V. W. EVELEIGH, *in reply*

I thank Dr. Eykhoff for his suggestion of an alternate technique and the example illustrating its use. Early in my study I searched for such an approach, but was led to the describing function method illustrated in the paper because of the flexibility it provides in handling various forms of  $IP$  surface characteristics. If the  $IP$  surface is such that the Laplace transform of eqn (4) in Dr. Eykhoff's discussion can be

readily taken, and it should often be reasonable to assume such a surface, then his suggested approach will work very nicely. In any particular situation, perhaps both approaches should be used to provide comparative results.

Regarding the filter  $G_2(s)$ , I agree with Dr. Eykhoff that it is not generally necessary nor even desirable to use such a filter to improve loop performance. The primary purpose for leaving it in was for flexibility in gathering experimental data on loop performance. It is in the low pass part of the loop and adjustments there result in direct changes in the equivalent low pass network, thereby requiring a minimum of effort to determine theoretical  $K_c$ ,  $\omega$  and limit cycle amplitude values (if such exist).

T. ISOBE, *University of Tokyo, Bunkyo -ku, Tokyo, Japan*

I highly appreciated Dr. Eveleigh's work which treats the problem of extreme searching adaptive systems by deriving an equivalent low pass filter and enables us to get a clear image of the solution. Incidentally a similar problem is dealt with, in the paper by T. Totani and myself, by using a sampled-data control system approach. I would like to ask Dr. Eveleigh a question. I imagine that the signal representing  $IP$  or  $f(x)$  is derived from a squared error or that sort of thing. How may we treat its fluctuation which is caused by randomness of the error of control?

V. W. EVELEIGH, *in reply*

I wish to thank Professor Isobe for his comments and question and also point out that I found his paper very interesting. I would suggest that a noise analysis should be possible using additive noise at the non-linearity or smoothing filter output with spectral characteristics chosen to be representative of a specific practical situation. Simulation results would be readily obtained and analytical results should be possible to obtain using reasonable assumptions. I hope to investigate this problem in further detail and report on the results at a later conference. For the examples chosen, the presence or absence of a reasonable level of noise does not appreciably alter the stability characteristics.



# The Realization of a Self-adapting Control Programme in a System with a Digital Computer

P. F. KLUBNIKIN

## Summary

The paper gives the results of a study of a system comprising a digital computer and a plant whose transfer function varies in time and is previously unknown. The plant and the digital computer are interlinked via transforming units. The system uses an adaptive control programme by which the coefficients of the plant transfer function are determined (an analogue of the plant is obtained). Control of the plant is effected in such a way that a criterion—the mean modulus of the error—is minimized. Determination of the plant transfer function coefficients is effected during the operation of the system.

The principle of operation of the control programme is as follows. In each cycle of operation of the system is performed an extrapolation of the input and search requirements (by use of 'fictitious' cycles within the machine) of the control signal which minimizes the system error.

The paper gives logical schemes of the control programme, as well as the results of an experimental investigation.

## Sommaire

Le rapport donne le résultat de l'étude d'un système comprenant un calculateur numérique électronique et une installation dont la fonction de transfert varie dans le temps et n'est pas connue à l'avance. Cette installation et ce calculateur sont reliés entre eux par des transmetteurs d'information. Ce système utilise un programme de commande adaptative simulant le système au moyen duquel les coefficients de sa fonction de transfert sont calculés. L'adaptation automatique de ce système consiste à minimiser le module moyen de l'erreur de réglage. La détermination des coefficients de la fonction de transfert du processus s'effectue lors du fonctionnement de celui-ci.

Le principe de fonctionnement du programme de ce système d'adaptation automatique est le suivant: dans chaque cycle d'opération du système, on effectue une extrapolation des valeurs d'entrée et du signal de commande nécessaire à la minimisation de l'écart de réglage, ceci grâce à des cycles fictifs internes au calculateur.

Le rapport donne le schéma fonctionnel du programme de commande ainsi que des résultats expérimentaux.

## Zusammenfassung

Der Aufsatz enthält die Untersuchungsergebnisse eines Systems, das einen Digitalrechner und eine unbekannte zeitveränderliche Regelstrecke enthält. Die Regelstrecke und der Digitalrechner sind über Umsetzer verbunden. Durch den Systemablauf, der einer selbststellenden Regelung entspricht, werden die Koeffizienten der Übertragungsfunktion der Regelstrecke bestimmt (man erhält ein Analogon der Regelstrecke). Als Kriterium der Regelung der Anlage dient das Minimum des Mittelwertes über den Betrag des Fehlers. Die Bestimmung der Koeffizienten der Übertragungsfunktion der Regelstrecke geschieht während des Regelablaufes des Systems.

Das Arbeitsprinzip des Regelvorganges ist wie folgt: In jeder Periode  $T$  des Arbeitsablaufes des Systems wird eine Extrapolation durchgeführt, die sowohl die Forderung an die Eingangsgröße als auch an die Selbststellung (die Rechenmaschine besitzt eine „fiktive“ Periode) bezogen auf das Regelsignal, das den Systemfehler zum Minimum macht, berücksichtigt.

Die Arbeit enthält Strukturschemen des Regelvorganges und experimentelle Untersuchungsergebnisse.

## Introduction

Recently there has been a wide expansion in systems in which the control of a load is achieved using a digital computer. In these systems by means of the application of the proper control programme DCM it is possible to obtain a self-adaptive (self-organizing) property, even in the case where there is no *a priori* information on all the properties of the load and the change of the load parameters in the course of time.

The elements of the theory of the construction of self-adaptive control systems are known, but their application in practice often results in very substantial difficulties connected with the special digital computing systems. The main difficulty is the determination of the characteristics of the load (for example, the transfer function) under conditions of normal operation of the system and the search for the control signal input to the load, which gives, from one or another point of view, the best control process.

This paper is devoted to the questions of the realization of a self-adaptive control programme in a system with a digital computer. One method of constructing a self-adaptive control programme is considered, which permits it to be realized relatively simply. Results of the experimental investigation of control systems are presented.

## The Method of Self-adaptation in the Control Programme

Consider an automatic control system consisting of a continuous part (the load) and a digital computer (DCM, *Figure 1*). The DCM operates in a realm of periodic repetition of a programme with a time cycle  $T$ .

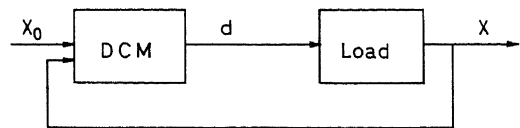


Figure 1. Diagram of the control system

Let the link between the control input  $X_0$  and the output quantity of the system  $X$  be given in the form

$$X^* = W_S(z) X_0^* \quad (1)$$

where  $X^*$  and  $X_0^*$  are the values of  $X$  and  $X_0$  at the moment of time  $T$ ;  $W_S(z)$  is the transfer function of the closed-loop system;  $z = e^{-q} = e^{-Ts}$  is a lag operator.

Then as is known<sup>1-3</sup>, in order for (1) to be satisfied the control system can be realized in the form of the block diagram shown in *Figure 2*, where for the transfer functions of the elements of the programme the following conditions should be satisfied

$$D_1(z) = W_S(z), D_3(z) = \frac{W_S(z)}{W_H(z)} \quad (2)$$

where  $W_H(z)$  is the discrete transfer function of the load. The transfer function  $D_2(z)$  is chosen arbitrarily. In the most simple case

$$D_2(z) = \frac{C_1 z + C_2}{C_3 z + 1} \quad (3)$$

However, it is impossible to satisfy the conditions (2) and, consequently to obtain the prescribed properties of the system, when the transfer function of the load  $W_H(z)$  is unknown or when its coefficients have an unknown time dependence, which is often the case in practice. It should be noted that in the indicated situation a general control programme, calculated in the presence of complete information about the load, is usually not convenient.

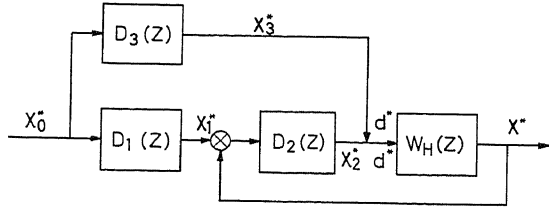


Figure 2. Diagram of the control programme

Therefore, the first step in a self-adaptive control programme is the determination of the discrete transfer function of the load, which is written in the form

$$\frac{X^*}{d^*} = W_H(z) = \frac{A_n z^n + A_{n-1} z^{n-1} + \dots + A_1 z}{B_n z^n + B_{n-1} z^{n-1} + \dots + B_1 z + B_0} \quad (4)$$

where  $n$  is the order of the load equations;  $A_i(t)$ ,  $B_j(t)$  are time-dependent coefficients ( $i = 1, 2, 3, \dots, n, j = 0, 1, 2, \dots, n$ ).

Consider that the computing-time cycle of the DCM is chosen so that the coefficients  $A_i(t)$  and  $B_j(t)$  are unable to change significantly over several cycles, and that  $n$  is unknown. Then in order to determine during the process of operation of the system the current values of the coefficients  $A_i$  and  $B_j$ , and consequently  $W_H(z)$  for a given moment of time, it is possible to use two simpler methods.

The first method is similar to that described in a previous work<sup>4</sup> and is based on the solution of a system of equations, which is obtained by using the expressions (4), i. e.

$$\begin{aligned} X_k B'_0 + X_{k+1} B'_1 + X_{k+2} B'_2 + \dots + X_{k+n} B'_n \\ = d_{k+1} A'_1 + d_{k+2} A'_2 + \dots + d_{k+n} A'_n \end{aligned} \quad (5)$$

where  $k = 0, 1, 2, \dots, 2n$

$$d_k = d(t - kT) u X_k = X(t - kT)$$

are the values of the input and output quantities of the load measured in the  $k$ th preceding calculation cycle;  $A'_i$  and  $B'_j$  are the approximate values of the coefficients for the current calculation cycle.

The system of eqns (5) is solved on the DCM relative to  $A'_i$  and  $B'_j$  by one of the known methods, for example by the method of iterations, and  $W_H(z)$  is determined thereby. The second method uses the principle of a 'learning model'<sup>5</sup> and includes the following.

Using the values of  $X_k$  and  $d_k$  ( $k = 0, 1, 2, \dots, n$ ) available in the memory of the DCM, a search is carried out by the

gradient method for the magnitudes of the coefficients  $A_i$ ,  $B_j$ , which give a minimum of the mean difference

$$\Delta_{av} = \frac{1}{m} \sum_{v=0}^m |X(t-vT) - X_M(t-vT)|$$

where  $\bar{m}$  is the number of cycles for averaging

$$X^* = W_H(z) d^*$$

$$X_M^* = W_{HM}(z) d^*$$

$W_{HM}(z)$  is a 'model' transfer function for the load formed in the DCM. This method is illustrated in Figure 3.

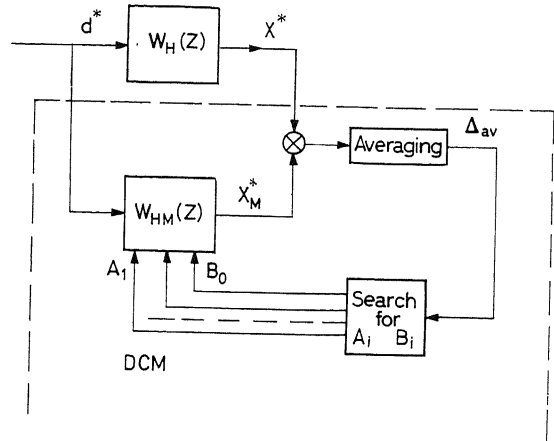


Figure 3. Diagram illustrating the method of determining the coefficients  $W_H(z)$ ,  $A_i$ ,  $B_j$

The second stage of the method described for building a self-adaptive control programme is the determination of a control signal  $d$ , which will guarantee the stability of the system and satisfy (1) or a better approximation to this condition. As a criterion for the approximation to (1) it is more useful to select the mean absolute value or the mean square of the error

$$\varepsilon_{av} = \frac{1}{\lambda} \sum_{v=0}^{\lambda} |\varepsilon(t-vT)| \quad (6)$$

$$\varepsilon_{av} = \frac{1}{\lambda} \sum_{v=0}^{\lambda} [\varepsilon(t-vT)]^2$$

where  $\lambda$  is the number of averaging cycles:

$$\varepsilon^* = [W_S(z) - W'_S(z)] X_0^*$$

$W'_S(z)$  is the transfer function of the closed-loop system,

$$W'_S(z) = \frac{W_H(z) [D_1(z) D_2(z) + D_3(z)]}{1 + D_2(z) W_H(z)} \quad (7)$$

Substituting in (7) the values of  $D_1(z)$  and  $D_2(z)$  from (2), one gets

$$W'_S(z) = \frac{\frac{W_H(z)}{W_{HM}(z)} + D_2(z) W_H(z)}{1 + D_2(z) W_H(z)} W_S(z)$$

Obviously in the general case  $W_{HM}(z) \neq W_H(z)$  and consequently  $W'_S(z) \neq W_S(z)$ . However, as experimental investiga-

tions have shown, even a relatively rough approximation  $W_{HM}(z)$  to  $W_H(z)$  for the condition of stability of the closed-loop circuit of the system (Figure 2), gives a behaviour of the system that is close to that prescribed.

Thus on the basis of (6) and (7) one has

$$\varepsilon_{av} = F(C_1, C_2, C_3, \dots) \quad (8)$$

The stability of the system is reached by seeking the minimum  $\varepsilon_{av}$  in the region of the coefficients  $C_1$ ,  $C_2$ , and  $C_3$ . The search is carried out by means of extrapolation in the DCM of  $X_0$  in  $r$ -conditional cycles and the calculation of  $\varepsilon_{avE}$  for these cycles.

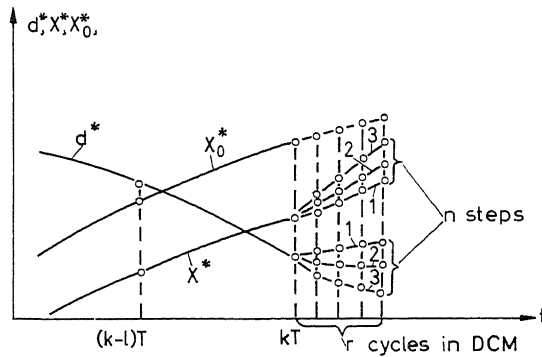


Figure 4. Graphs illustrating the method of constructing a self-adaptive programme

The idea of the method is explained in the diagrams shown in Figure 4. As a result, for each cycle of the DCM the following order of operations is obtained:

- (1) Input  $X_0$  and  $X$ .
- (2) Extrapolation of  $X_0$  in  $r$ -conditional cycles.

In the simplest case for linear extrapolation from the preceding cycle one gets

$$X_{0E}(t+rT) = r[X_0(t) - X_0(t-T)] + X_0(t) \quad (9)$$

- (3) Determination of the coefficients of  $W_H(z)$ .

(4) The search for the minimum  $\varepsilon_{avE}$  in the region of the coefficients  $D_2(z)$  taking into account the next  $r$  cycles.

For the method of the modified gradient, on the basis of (8), one has the following formulae

$$\varepsilon_{avE}^{(C_1)} = \frac{F_\Delta(C_1 + \Delta C, C_2, C_3) - F(C_1, C_2, C_3)}{\Delta C}$$

$$\varepsilon_{avE}^{(C_2)} = \frac{F_\Delta(C_1, C_2 + \Delta C, C_3) - F(C_1, C_2, C_3)}{\Delta C}$$

$$\varepsilon_{avE}^{(C_3)} = \frac{F_\Delta(C_1, C_2, C_3 + \Delta C) - F(C_1, C_2, C_3)}{\Delta C}$$

$$\Delta C_1 = -k\varepsilon_{avE}^{(C_1)}$$

$$\Delta C_2 = -k\varepsilon_{avE}^{(C_2)}$$

$$\Delta C_3 = -k\varepsilon_{avE}^{(C_3)}$$

$$D'_2(z) = \frac{(C_1 + \Delta C_1)z + C_2 + \Delta C_2}{(C_3 + \Delta C_3)z + 1}$$

$$X_{1E}^* = W_S(z) X_{0E}^*, \quad X_{2E}^* = (X_{1E}^* - X_E^*) D'_2(z)$$

$$X_{3E}^* = \frac{W_S(z)}{W_{HM}(z)} X_{0E}^*, \quad X_E^* = W_{HM}(z) d_E^*$$

$$d_E^* = X_{2E}^* + X_{3E}^* \quad (10)$$

[Note: the quantity  $\Delta C$  can be taken equal to unity.]

where  $\varepsilon_{avE}^{(C)}$  are the partial derivatives with respect to the coefficients  $D_2(z)$ ;  $k$  is the coefficient of a step in the direction of the reversed gradient;  $\Delta C$  is the trial increment;  $X_E$ ,  $X_{1E}$ ,  $X_{2E}$ ,  $X_{3E}$ ,  $d_E$  are the values of the corresponding quantities in the conditional cycles within the DCM (the index  $E$  indicates extrapolated values of the corresponding quantities).

(5) After an  $m$  step search the output signal from the DCM  $d$  is calculated in accordance with the diagram shown in Figure 2.

$$d(t) = X_2(t) + X_3(t)$$

$$X_2(t) = C_2[X_1(t) - X(t)] + C_1[X_1(t-T) - X(t-T)]$$

$$- C_3 X_2(t-T)$$

$$X_1(t) = G_1 X_0(t-T) + G_2 X_0(t-2T) + \dots + G_m X_0(t-mT)$$

$$X_3(t) = \frac{1}{A_1} G_1 X_0(t) + \frac{1}{A_1} (G_1 B_1 + G_2) X_0(t-T)$$

$$+ \dots + \frac{1}{A_1} B_n G_m X_0[t - (m+n-1)T] - \frac{A_2}{A_1} X_3(t-T)$$

$$- \dots - \frac{A_n}{A_1} X_3[t - (n-1)T] \quad (11)$$

Here one takes

$$W_S(z) = G_1 z + G_2 z^2 + G_3 z^3 + \dots + G_m z^m$$

(6) The output of the control signal  $d$  and the updating of the information in memory.

In Figure 5 is shown the logical flow diagram of a self-adaptive DCM programme which assures that the given operations and the calculations according to the formulae (9)–(11) will be carried out. Circles indicate conditional transfer operators (transfer control), and the conditions are written inside them. As can be seen from the flow chart, 15 cycles are provided for in the control programme, in the course of which normal control is achieved according to the diagram shown in Figure 2. This is required for the accumulation of information in the DCM. No particular explanation is required for the remainder of the flow chart.

It must only be noted that the number of conditional cycles in the DCM must be chosen so that the time for accomplishing the operations described does not take longer than the cycle time  $T$ . If, for a minimum number of conditional cycles (one or two), it is not possible to satisfy this condition, then it is necessary to use a DCM that is faster acting (in which each arithmetical or logical operation is executed in less time).

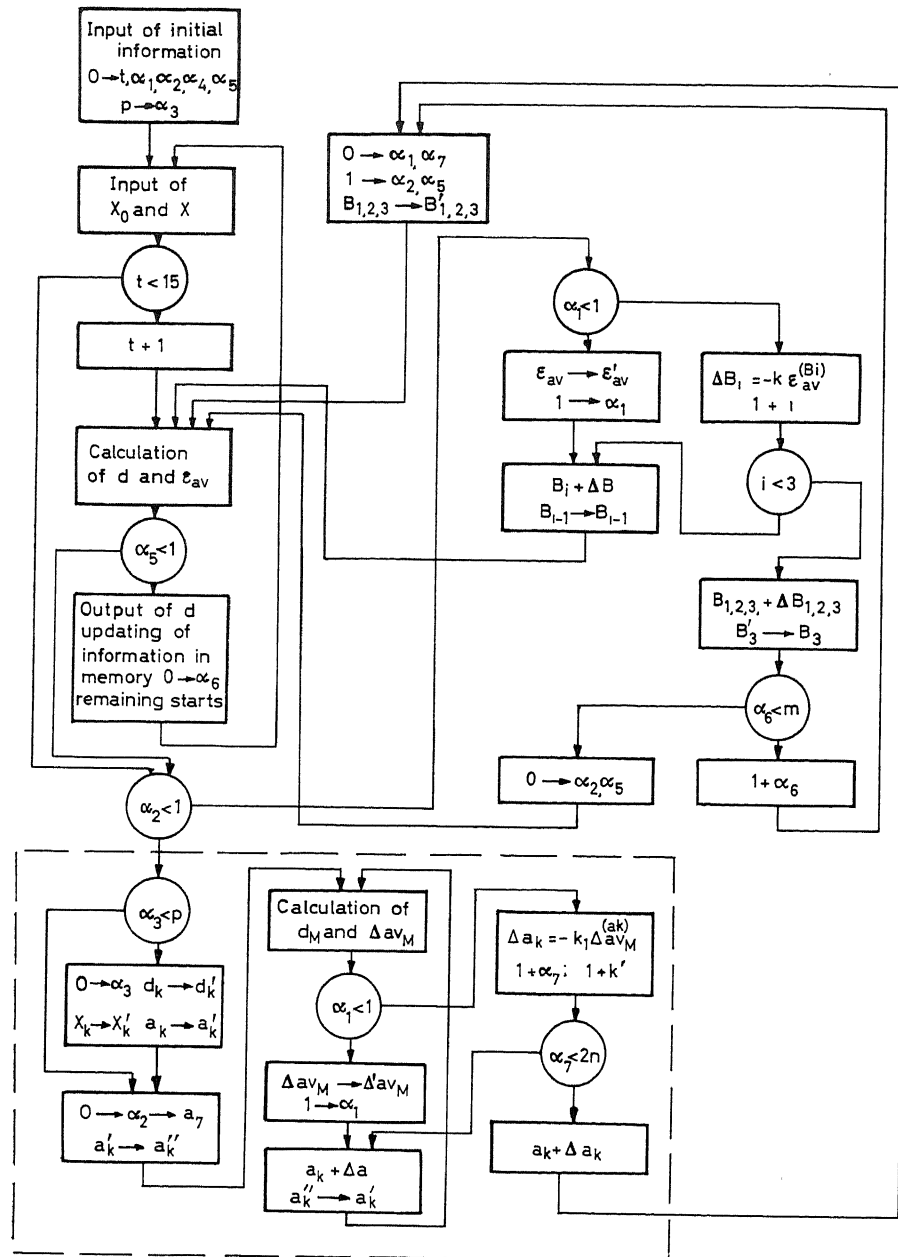


Figure 5. Flow chart of the control programme  
 $p$ -number ( $p > k$ ) defining input condition of information in  $X_k, d_k$

### Results of Experimental Investigations

In carrying out the experimental investigation of the load in the system of Figure 1 its dynamic model was changed. The dynamic model of the load was linked with the DCM through a device transforming a voltage into an 8-digit binary code or the code into a voltage. The control input  $X_0$  is supplied in the form of a voltage and fed through the transforming device to the DCM. The diagram for the realization of the control system during the performance of the experiment is shown in Figure 6.

A control programme was fed into the DCM corresponding to the flow diagram shown in Figure 5. The dynamic model of the load was characterized by the transfer function

$$W_H(S) = \frac{K(T_0 S + 1)}{S(T_1^2 S^2 + 2T_1 \xi S + 1)} \quad (12)$$

The quantities  $K$ ,  $T_0$ ,  $T_1$ , and  $\xi$  can be varied in time over the following limits:  $K = 0.1 \div 0.001$ ;  $T_0 = 1.5 \div 0.2$ ;  $T_1 = 0.2 \div 0.5$ ;  $\xi = 0.2 \div 0.05$ .

The rate of change of the quantities indicated did not exceed 1–5 per cent/sec from the initial value. The calculation cycle in the DCM was equal to  $T \cong 0.15$  sec. The connection between the control input  $X_0$  and the output of the system  $X$  was given in the form

$$W_S(z) = \frac{1}{3}(z + z^2 + z^3) \quad (13)$$

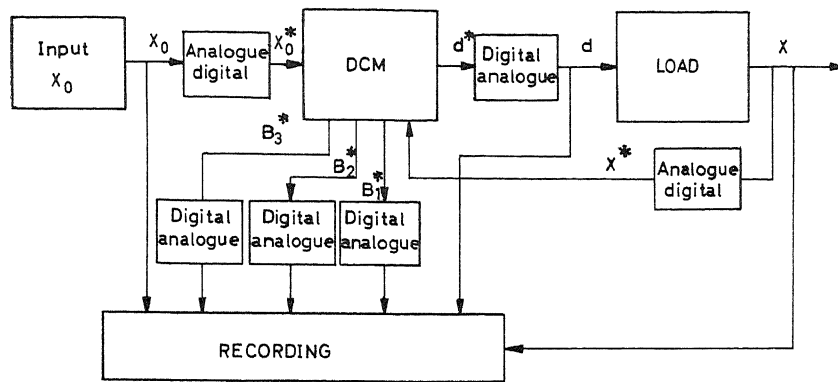


Figure 6. Diagram of control system for performing experiment  
Analogue-digital transforms voltage to binary code. Digital-analogue transforms binary code to voltage

For fixed values of the coefficients  $W_H(S)$  of (12) and with fulfilment of the conditions (2) the system has a first-order instability and a transfer process defined by (13). The rate gain in the system is relatively small. Its increase is limited by an

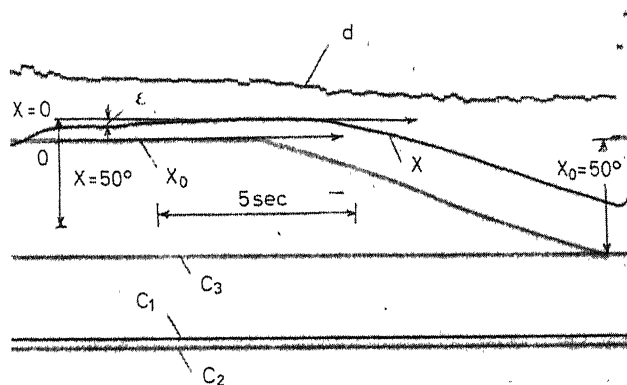


Figure 7. Oscillogram of the evolution of an input control system with constant coefficients  $D_2(z)$  with normal control  
 $C_1 = 0.25$ ,  $C_2 = 0.5$ ,  $C_3 = -0.25$  selected by numerical means

instability in the closed-loop circuit for the selected structure  $D_2(z)$  of (3).

In Figure 7 is shown an oscillogram for the operation of a control system with normal control (self-adaptive programme excluded). As the experiment shows, for normal control the system is extremely sensitive to a change of the coefficients  $D_2(z)$ , especially when this leads to an increase in the gain of the closed-loop circuit. In this case a change in the coefficients  $C_1$ ,  $C_2$ , and  $C_3$  by 10–15 per cent makes the system unstable. The same effect occurs in the system with a change in  $W_H(z)$ .

By putting a self-adaptive control system into operation in a short time (10–15 sec) the optimal value of the coefficients  $D_2(z)$  was found and the error was reduced to a minimum. In the process of operating, the system automatically adapted itself to the changed characteristics of the load.

In Figure 8 are shown typical curves of the change of the coefficients  $W_H(z)$  during their determination in the DCM. The transfer function corresponding to (12) is written in the form

$$W_H(z) = \frac{a_1 z + a_2 z^2 + a_3 z^3}{(1-z)(a_4 z^2 + a_5 z + 1)} \quad (14)$$

It can be seen from the curves that even in 6–8 sec the coefficients of (14)  $a_i$  approximate their true values indicated on the graph by broken lines.

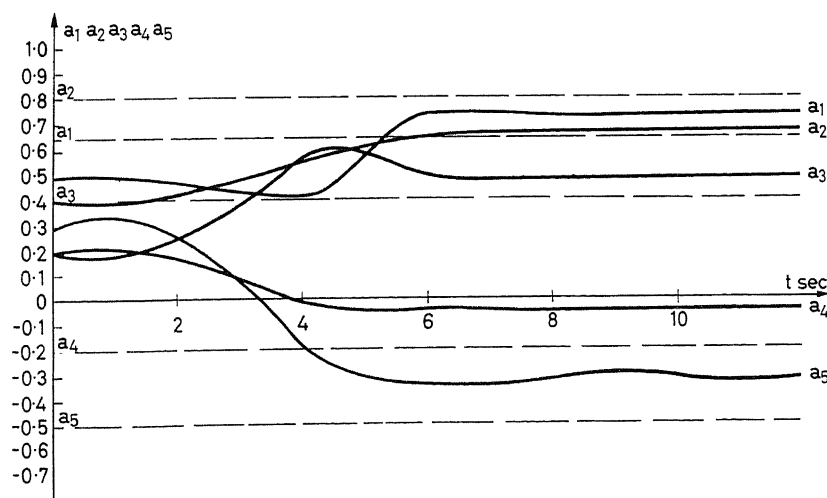


Figure 8. Graph of the change of the coefficients  $a_i$  in the process of searching in the determination of  $W_{HM}(z)$

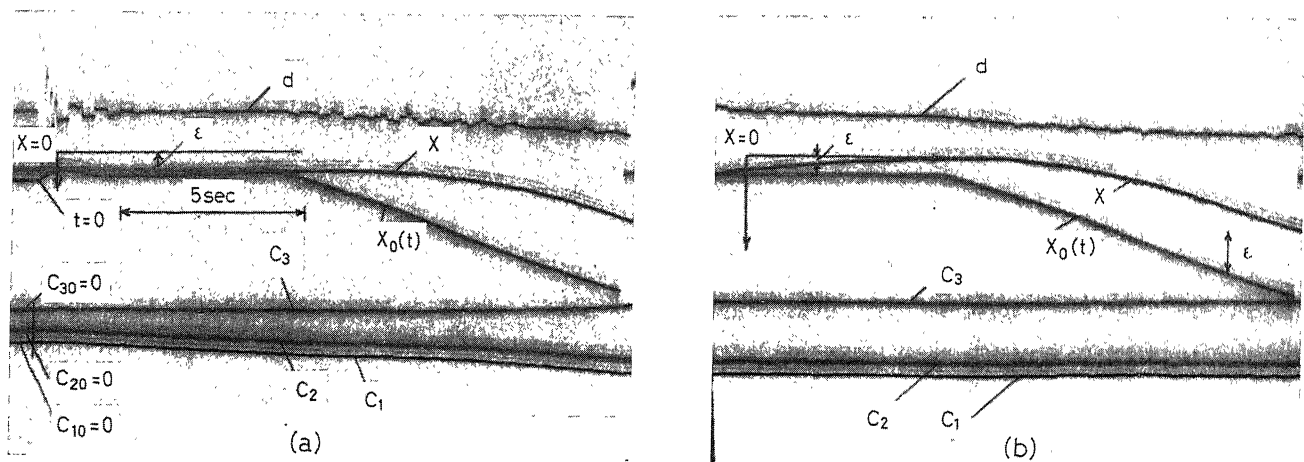


Figure 9. Oscillograms of the operation of the system.  
 $\Delta C = 2^{-5}$ ,  $K = 0.98$ ,  $m = 3$ ,  $r = 2$ ,  $\lambda = 1$   $C_{10} = C_{20} = C_{30} = 0$   
 (a) portion of initial operation  
 (b) portion after forming the coefficients

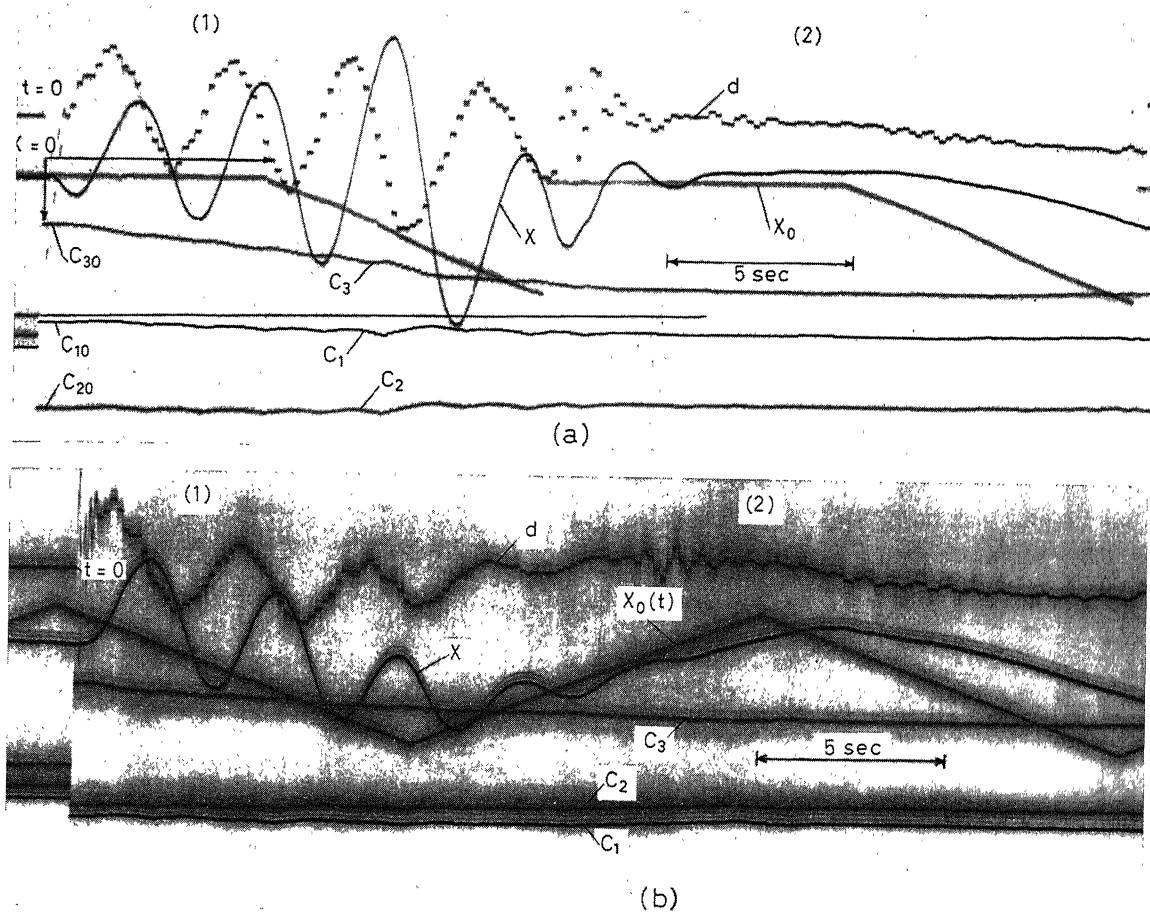


Figure 10. Oscillograms of the operation of the systems:  $C = 2^{-5}$ ,  
 $K = 0.98$ ,  $m = 3$ ,  $r = 2$ ,  $\lambda = 1$   
 (1) initial operation  
 (2) after forming the coefficients  
 (a)  $C_{10} = -0.98$   $C_{20} = -0.25$   $C_{30} = 0.8$   
 (b)  $C_{10} = -0.9$   $C_{20} = 0.22$   $C_{30} = 0.23$

The curves in Figure 8 were made for the very worst case, where the determination is carried out by a step input to the system applied at the time  $t = 0$ , after which  $X_0$  remains constant. For an arbitrary time change of  $X_0(t)$  the errors in the determination of the coefficients are significantly decreased and do not exceed 5–10 per cent.

In Figures 9 and 10 are shown oscillograms showing the evolution of the system in the process of changing the coefficients  $D_2(z)$ ,  $C_1$ ,  $C_2$ , and  $C_3$ . The oscillograms in Figure 9 correspond to a combination of initial values of  $C_1$ ,  $C_2$ , and  $C_3$  for which the total gain of the closed-loop circuit of the system, i.e.,  $W_H(1) D_2(1)$  is small and for which the variable input signal  $X_0(t)$  error is large. The oscillograms in Figure 10 correspond to initial values  $C_1$ ,  $C_2$ , and  $C_3$ , which lead to an unstable system. In both cases, in a relatively small time the system automatically selects the optimum value of the coefficients  $D_2(z)$ , for which the error is a minimum for the given control input  $X_0(t)$ .

[It is interesting to note that when the load simulator is switched off ( $X = \text{const.}$ ) in the course of a few cycles of the operation of the DCM the quantity  $\varepsilon_{avB}$  is reduced to a minimum and also for  $X_0(t)$ , which indicates the efficiency of the search method used.]

## Conclusions

The proposed method of constructing a self-adaptive control programme can easily be realized in a DCM and requires a relatively small number of instructions in the programme.

For the determination of the dynamic properties of the load, in the process of normal operation of the automatic control system with a DCM, it is useful to use a transfer function in the form  $W_H(z)$  (the equivalent of a difference equation). Here a good result in the determination of the coefficients of  $W_H(z)$  gives a method presented above, which is based on the principle of a 'learning model'.

The experimental investigation showed the efficiency of the self-adaptive control programme, constructed according to the proposed method.

## References

- 1 TSYPKIN, YA. Z. Theory of pulse systems. *Fizmatgiz* (1958)
- 2 TOU, J. *Digital and Sampled-data Control Systems*. 1959. New York
- 3 KLUBNIKIN, P. F. Synthesis of control programmes in systems including digital calculating machines. *Automat. Telemekh.* Vol. 21, No. 11 (1960)
- 4 KALMAN, R. E. Planned self-organizing control systems. *Trans. Amer. Soc. mech. Engrs* No. 57 (1957)
- 5 WIDROW, B. Adaptive sampled-data systems. *Automatic and Remote Control*. 1960. London; Butterworths

## DISCUSSION

A. R. M. NOTON, *Electrical Engineering Dept., University of Nottingham, Nottingham, England*

My question relates to the identification problem. In order to evaluate the coefficients of eqn (4) the author mentions two possible approaches: (a) that of eqn (5) and (b) the so-called learning model. Will he discuss the relative merits of the two approaches, explaining why he has concentrated on the second method?

H. H. ROBERTSON, *I.C.I. Ltd., Wilton Works, Middlesbrough, Yorkshire, England*

In the author's discussion of the determination of the process transfer function, he considers a form given by eqn (4). This may be written

$$W_H(z) = \frac{\sum_{i=1}^N A_i z^i}{\sum_{i=0}^n B_i z^i}$$

An alternative form for this is given by

$$W_H(z) = \frac{Q(z)}{P(z)} = \frac{A \prod_i (z - s_i)}{\prod_i (z - r_i)}$$

where  $s_i$  and  $r_i$  are the zeros and poles respectively of the transfer function. It is important to choose the number of poles in the transfer function to correspond with the order of the system.

The behaviour of the system can be expressed as

$$X_m^* = \sum_i c_i r_i^m + \text{output due to input.}$$

If the model transfer function has more poles than the actual system, then spurious poles have been introduced which may cause instability if  $|r| > 1$ . Moreover the least squares equation corresponding to (5) is singular and it is very likely that numerical determination of  $A_i$  will produce a pole with  $|r| > 1$ .

If no poles are taken in the model transfer function and a sufficiently large value of  $n$  is used, we obtain the unit impulse response function in sampled data form. Work on this problem has taken place at I.C.I. and Figure A shows a typical kernel of coefficients  $A_i$ . This kernel was produced from data taken from an industrial plant and it was found necessary to pass the time series through a high-pass filter in order to eliminate the effects of slow drift in one of the variables.

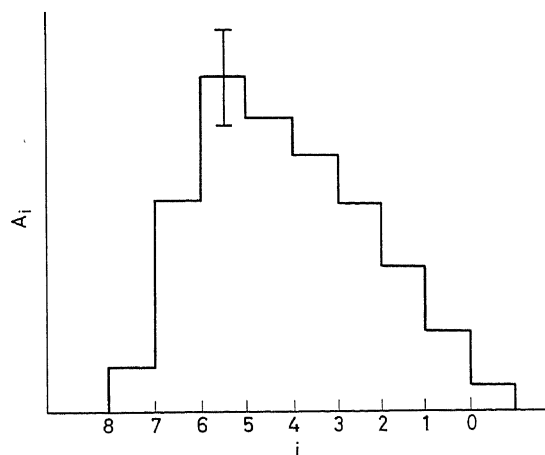


Figure A

P. F. KLUBNIKIN, *in reply*

In reply to Dr. Noton, the second method of obtaining transfer function (4) in the digital computer was adopted because it was more effective.

It was not always possible to obtain a solution of system (5) when the differences  $d_k \dots d_{k+n}$  were very small. Moreover, the solution required more time in computer cycles.

The author thanks Dr. Robertson for an interesting discussion.

# A Pattern Recognizing Adaptive Controller

W. K. TAYLOR

## Summary

When a process is controlled by a number of parameters it may be possible for an optimizing controller or hill climber to seek and find optimum parameter values by trial and error parameter perturbation techniques. The general effect of uncontrolled disturbances is to change continually the state of the process, and the parameter values are required to track the changing optimum operating conditions. The accuracy of the tracking is limited by the relatively slow response of the optimizing controller since the process must be allowed time to respond to repeated perturbations of all the parameters. A method of steadily improving the tracking, and hence the average operating efficiency of a process, by means of a learning system has been developed. The learning system automatically stores information about the process states and the associated optimum parameter values. This information is utilized subsequently to recognize process state patterns and thereby to select the appropriate optimum parameter values with negligible delay, since the parameter optimizer is no longer involved.

The paper gives the theory of the method and shows how patterns that are characteristic of the process state automatically control first the storage and subsequently the selection of the optimum parameter values. A practical pattern recognizing controller (p. p. r. c.), designed according to the theory and employing transistor circuits, is described in detail. Preliminary tests, using a small number of inputs, have shown that the controller behaves as predicted by the theory.

## Sommaire

Dans la commande optimalisante d'une installation à paramètres multiples où les conditions optimales varient en raison des perturbations incontrôlables, les valeurs optimales des paramètres peuvent être trouvées par des tâtonnements successifs. La précision de l'optimaliseur est toutefois limitée par le temps nécessaire aux tâtonnements. Un système d'apprentissage, ne présentant pas cet inconvénient a été développé. Ce système mémorise automatiquement les états de fonctionnement de l'installation ainsi que les valeurs optimales associées de ses paramètres. Ces renseignements sont ensuite utilisés pour éléctionner les valeurs optimales des paramètres par simple identification de la structure de l'état de fonctionnement. Cette sélection n'implique pas l'utilisation de l'optimaliseur, donc, elle dure très peu de temps.

On décrit la théorie de la méthode et on montre comment les structures caractéristiques de l'état de l'installation commandent automatiquement d'abord la mémorisation, et ensuite, la sélection des valeurs optimales des paramètres. On décrit les détails d'un adaptateur par identification de structure réalisé d'après ce principe et utilisant des circuits à transistor. Les premiers essais, avec un nombre restreint d'entrées, confirment l'analyse théorique.

## Zusammenfassung

Wird eine Strecke durch eine Anzahl von Parametern gesteuert, so ist es einem selbstanpassenden Regler möglich, die günstigsten Reglereinstellungen durch versuchsweise Störung der Parameter zu suchen und zu finden. Im allgemeinen führen unregelmäßige Störungen zu einem dauernden Zustandswechsel der Strecke und die Einstellwerte müssen den wechselnden günstigsten Betriebsbedingungen folgen. Die Nachlaufgenauigkeit wird durch das relativ langsame Verhalten des anpassenden Reglers begrenzt, da die Strecke Zeit braucht, um auf wiederholte Störung aller Kennwerte zu reagieren. Ein Verfahren zur stetigen Verbesserung des Nachlaufens und damit der mittleren

Leistung der Strecke wurde aufgrund eines Lernsystems entwickelt. Dieses speichert selbsttätig Informationen über den Zustand der Strecke und die zugehörigen günstigsten Einstellwerte. Diese Information wird später dazu benutzt, Muster des Streckenzustandes zu erkennen und dadurch die günstigsten Einstellungen mit geringster Verzögerung zu wählen, weil ja ein Optimiergerät nicht mehr benötigt wird.

Der Beitrag gibt die Theorie der Methode an und zeigt, wie die Muster des Streckenzustandes zuerst die Speicherung und dann die Auswahl der günstigsten Einstellwerte regeln. Eine technische Ausführung des mustererkennenden anpassenden Reglers, der gemäß der hier entwickelten Theorie gebaut wurde und Transistorschaltkreise benutzt, wird in Einzelheiten beschrieben.

Vorläufige Ergebnisse mit einer geringen Anzahl von Eingängen zeigen gute Übereinstimmung mit der Theorie.

## Introduction

The optimization of a complex process by means of optimizing controllers employing trial and error parameter perturbation techniques has a time course that is long compared with the process response time, since it must in general be based on many small parameter changes and the process must be allowed time to respond to each change. This may be the best that can be achieved when the parameter optimizing controller is first used, but if appropriate information about the state of the process and the optimum parameter values is selected and automatically stored in a permanent but adaptive memory system it is possible for the average optimization response time to be decreased as the store of information grows. Thus, if the state of the process can be adequately defined by taking a sufficiently large number of measurements, including the optimum parameter values, when an optimum state is finally reached, it is only necessary to recognize this state immediately whenever it occurs on future occasions and to use this information to select the associated optimum parameter values at the same instant. By this means the long initial optimization delay in response to a change in the process state, which may be over an hour for some processes, can be completely eliminated for the particular state that can be recognized and also for process states that so resemble it that they are similarly classified by the pattern recognizer.

In practice the state of a process and the corresponding optimum parameters are continually changing due to uncontrolled variables and 100 measurements may be required for an adequate specification of the state. For a 10 per cent accuracy of measurement the theoretical number of possible distinct states is  $10^{100}$  and in general each may require a different set of parameter values for optimum performance. In practice, however, there are usually severe physical constraints on the process states and on the factors that cause the state to change so that there may be a possibility of obtaining worthwhile increases of efficiency with a pattern recognition system of reasonable memory capacity.



### Operating Principle of the Pattern Recognizing Adaptive Controller

In *Figure 1* the inputs to the pattern recognizing adaptive controller (p.r.a.c.) derived from the controlled process consist of  $N$  measurements  $i_1 \dots i_N$ , and to illustrate the operation of the controller  $N$  will be assumed to be 100, as in the experimental system that has been constructed. The measurements may, for example, indicate the distribution of temperature throughout a vessel and in this case there will be a physical restriction on the temperature gradient and typical set of readings may be as shown in *Figure 2 (a)*, curve *A*. This is in fact a sample of band limited random noise obtained from a noise generator and could represent deviations of the measurements from reference levels. If the measurements are not sufficient to describe the state of the process adequately, they may be extended to include the distributions of other variables such as pressure and specific gravity.

It is assumed that two parameters  $P_1$  and  $P_2$ , each having ten possible values, are controlled and that an optimum setting of these exists for the process state corresponding to the 100 measurements with the envelope *A* in *Figure 2 (a)*. At time  $t_1$  the optimum values of the two parameters denoted by  $P_{10}(t_1)$  and  $P_{20}(t_1)$ , will correspond to a maximum of the measure of performance such as  $M_2$  in *Figure 2 (b)* and, if  $M_2$  remains stationary or moves sufficiently slowly,  $P_1$  and  $P_2$  will eventually reach their optimum values under the control of the parameter optimizer. Any disturbances that are sufficient to change the optimum operating point to  $M_4$  at  $t_2$ , for which the optimum parameter values are  $P_{10}(t_2)$  and  $P_{20}(t_2)$ , can also be expected to change the state of the process and hence the distribution of measurements as illustrated by curve *B* in *Figure 2 (a)*. For sudden disturbances the curve will change from *A* to *B* at a rate determined by the process time constants, whereas the parameters will change at a much slower rate due to the optimizer

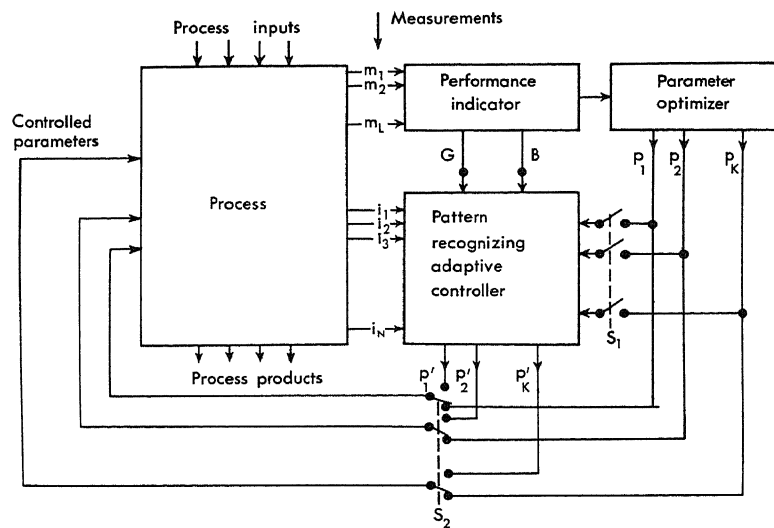


Figure 1

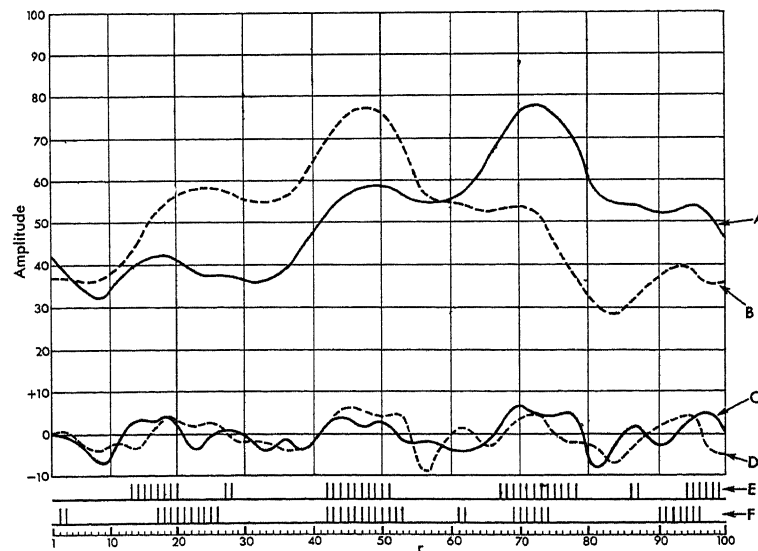


Figure 2 (a)

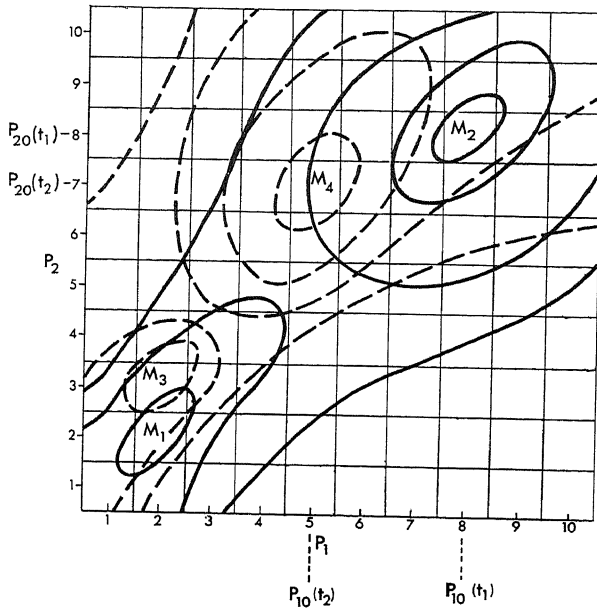


Figure 2 (b)

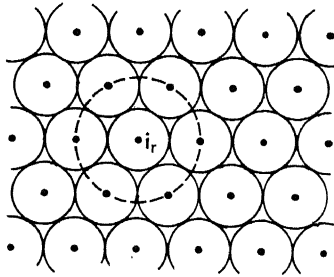


Figure 2 (c)

time constants. This optimizer lag is entirely eliminated when the p.r.a.c. is able to recognize the measurement distributions and select the optimum parameter values instantaneously on the basis of past information. The accumulation of this information will generally be a very slow process but it may be speeded up by introducing artificial disturbances so that a wide range of sample process states are covered in the shortest time that will allow the optimizer to track the moving optimum point.

The principal problems considered in this paper are the theory of operation and design of a p.r.a.c. that will learn to recognize the measurement pattern and to supply the most probable optimum parameter values on the basis of earlier experience. This result could clearly be achieved in principle by storing the information in a digital computer of sufficient storage capacity but the method described here leads to a low cost, on-line p.r.a.c. employing a parallel memory system giving negligible access time to the entire memory and extreme reliability. The controller is thus suitable for high speed processes for which a digital computer would be far too slow.

#### Detection of Non-linear Changes in the Measurement Pattern

The first stage in the p.r.a.c. simplifies the input measurements  $i_1 \dots i_N$  by forming a new set of variables  $j_1 \dots j_N$  defined by

$$j_r = K \left[ i_r - \frac{1}{2} (i_{r+q} + i_{r-q}) \right] \quad (1)$$

in which the constant and linear changes in the original envelope are essentially eliminated. This operation is performed in unit  $H$  (Figure 3) by inverting the  $i_r$  and forming the appropriate combinations in resistor networks. The  $j_r$  corresponding to curves  $A$  and  $B$  in Figure 2 (a), are shown respectively at  $C$  and  $D$  for the case when  $q = 3$  and  $K = 3$  in eqn (1).

In some processes it may be necessary to take measurements over a two or three-dimensional field and in these cases eqn (1) is extended to include signals surrounding each  $i_r$ . A two-dimensional  $H$  unit has been constructed for the experimental p.r.a.c. and this operates on 100 measurements taken over a  $10 \times 10$  plane of transducers arranged in a hexagonal pattern as shown in Figure 2 (c). The equation corresponding to this arrangement is

$$j_r = K \left[ i_r - \frac{1}{6} (\text{Sum of six measurements on circle surrounding } i_r) \right] \quad (2)$$

The effect of the  $H$  unit in one or more dimensions can best be illustrated by the curves  $C$  and  $D$  of the one-dimensional example which tend to have peaks in the regions where the slope of the original curves change. Each output  $j_r$  of the  $H$  unit is supplied to an  $I$  unit and also to an  $A_1$  and  $B_1$  unit. The latter replace the  $j$  signals that are above or below the average level by constant amplitude signals  $O_{B_1}$  and  $O_{B_2}$  respectively. The  $O_{B_1}$  signals for the two measurement patterns are shown at  $E$  and  $F$  in Figure 2 (a).

The next sections in the p.r.a.c. learn associations between the  $j$  patterns and the optimum parameter values on a maximum likelihood basis so that the parameters selected by any new  $j$  pattern are initially the same as those selected by the nearest learnt pattern, being the most probable values. If, however, a new pattern is associated with different parameters during the learning process, then these parameters become the most probable values for the new pattern which becomes one of the learnt patterns. In other words, the p.r.a.c. makes the best guess on the evidence currently available and the high speed at which this is done is due to the parallel method of storing this evidence and of obtaining simultaneous overall access to it, as described in the next section.

#### Theoretical Basis of the Learning and Recognition Process

The  $N$  outputs  $j_1 \dots j_N$  of the  $H$  unit are supplied through diodes to the potentiometers of the information storage units  $I$ , as shown in Figures 3 and 4(d), the diodes suppressing negative values. All the potentiometer sliders are initially at the lower end so that the voltages at the sliders are initially one half of the positive inputs or zero in the case of negative inputs. The sliders may be driven independently by the ratchet impulse motors in the direction that increases the gain ' $a$ ' of the  $I$  unit from the initial value  $\frac{1}{2}$  towards a final value unity that is not normally reached. The  $I$  units are grouped first into pairs and then into  $M$  sections of  $N$  pairs,  $M$  being the number of quantized levels for each parameter. The outputs of the right-hand  $I$  units,  $a_r \cdot j_r$  in each pair, are passed through a sign changing summing amplifier  $A_2$  that forms

$$-\frac{1}{N} \sum_{r=1}^N a_r \cdot j_r \quad \left( \frac{1}{2} \leq a_r \leq 1 \right) \quad (3)$$

This quantity is supplied to the summing integrator  $D$  together with the left-hand  $I$  unit outputs  $a_{r+} \cdot j_r$  that are not changed in sign. The gain of the summing integrator is adjusted so that its output becomes

where  $C_1$  and  $C_2$  are constants and  $t = 0$  is the time at which a reset pulse, that has reset the integrator output to zero, is removed. The gains  $a_{r+}$  and  $a_{r-}$  are initially  $\frac{1}{2}$  throughout the p.r.a.c.  $I$  units and the integrator outputs increase linearly in magnitude after a reset pulse until one of the  $B_2$  units is triggered. Theoretically, the  $B_2$  unit that triggers first is initially selected at random, but in practice it is found that small errors in component values cause a bias in favour of certain units.

$$\begin{aligned} & - \int_0^t \left[ C_1 + C_2 \left\{ \sum_{r=1}^n \left( \frac{1}{2} + \frac{d}{2n} \frac{1}{2} \right) j_r \right. \right. \\ & \left. \left. + \sum_{n+1}^N \left( \frac{1}{2} - \frac{1}{2} - \frac{d}{2n} \right) j_r \right\} \right] dt \\ & = - \int_0^t \left[ C_1 + C_2 \sum_{r=1}^n \frac{d}{2n} j_r \right] dt \end{aligned} \quad (5)$$
$$-\int_0^t \left[ C_1 + C_2 \left\{ \sum_{r=1}^n \left( \frac{1}{2} + \frac{d}{2n} - \frac{1}{2} - \frac{d}{2n} \right) j_r \right. \right. \\ \left. \left. + \sum_{n+1}^N \left( \frac{1}{2} + \frac{d}{2n} - \frac{1}{2} - \frac{d}{2n} \right) j_r \right\} \right] dt = \int_0^t C_1 dt$$

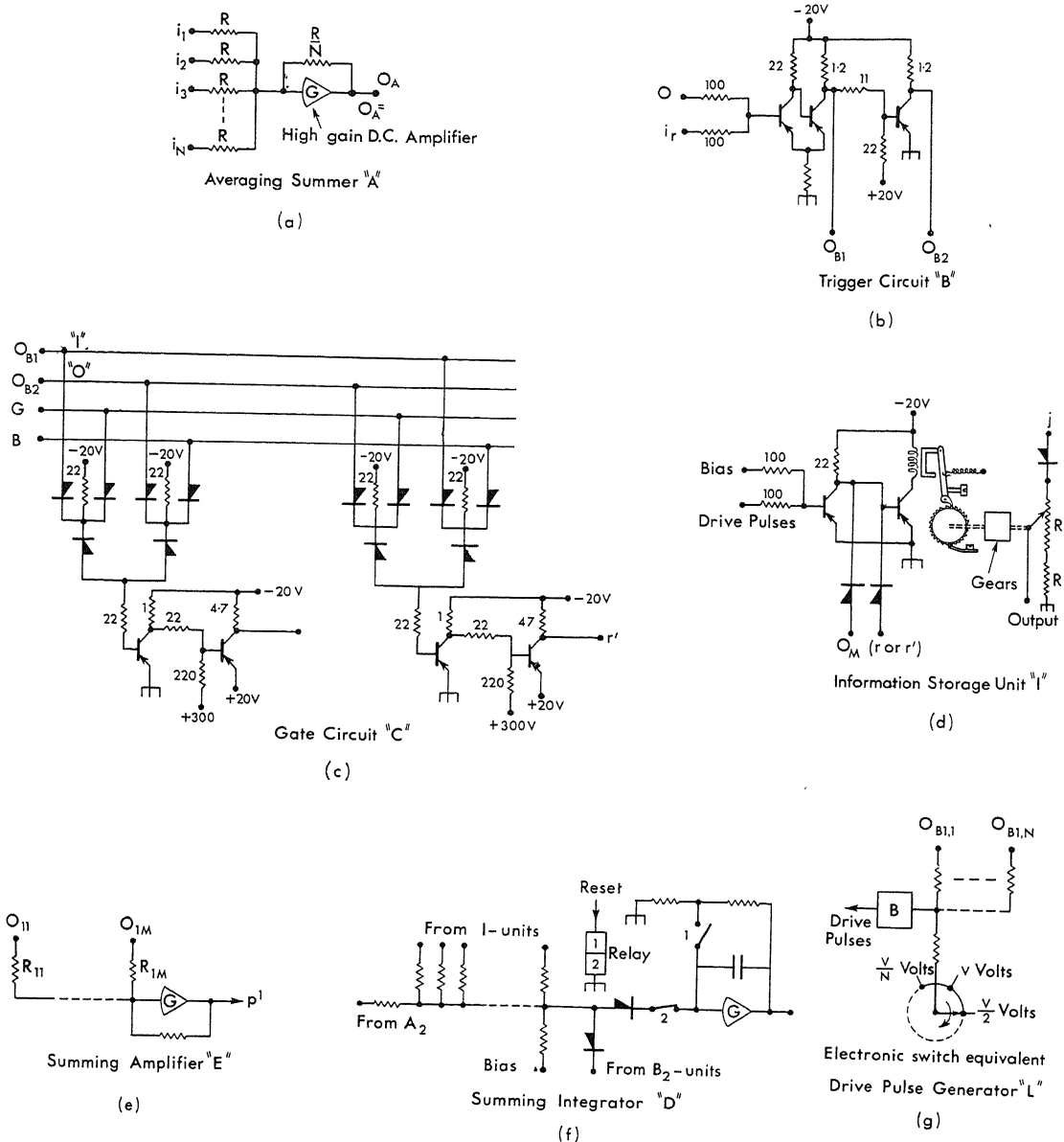


Figure 4(a-g)

This facility can be used to cancel the effect of any changes that are subsequently found to be in error or inefficient as, for example, when the parameters supplied initially correspond to a secondary lower maximum such as  $M_1$  or  $M_3$  in Figure 2 (b).

The next step in the theory will be to show that any chosen pattern of  $n$  ( $1 \leq n \leq N$ ) equal non-zero signals of any amplitude and  $N - n$  zero signals will cause the  $B_2$  unit supplied by  $I$  units with  $n$  corresponding gains  $a_r$  at  $1/2 + d/2n$  and  $(N - n)$  corresponding  $a_{r-}$  gains at  $1/2 + d/2n$  to trigger before any other  $B_2$  unit, after a reset pulse. The closest possible sets of  $I$  unit gains to the particular set chosen are the two sets that have only one difference, namely the set with  $(n - 1)$  gains  $a_{r+}$  at  $1/2 + d/2n$  and  $(N - n + 1)$  gains  $a_{r-}$  at  $1/2 + d/2n$  and the set with  $(n + 1)$  gains  $a_{r+}$  at  $1/2 + d/2n$  and  $(N - n - 1)$  gains  $a_{r-}$  at  $1/2 + d/2n$ . Let the contributions of these two sets

to the integrands be  $q_{n-1}$  and  $q_{n+1}$  respectively and let the contribution due to the chosen pattern be  $q_n$ . When the chosen pattern is present at the p.r.a.c. input terminals the value of  $q_n$ , as given in eqn (5), is

$$q_n = \frac{d}{2n} \sum_{j=1}^n j_r = \frac{dj}{2} \quad (6)$$

since the  $n -$  non-zero  $j_r$  are assumed to be equal. The values of  $q_{n+1}$  and  $q_{n-1}$  for the same input pattern are

$$q_{n-1} = \sum_{j=1}^{n-1} \left[ \left( \frac{1}{2} + \frac{d}{2(n-1)} - \frac{1}{2} \right) j_r + \left( \frac{1}{2} - \frac{1}{2} - \frac{d}{2(n-1)} \right) \right] j_r \\ = \frac{(n-2) dj}{(n-1) 2} \quad (7)$$

$$q_{n+1} = \sum_1^n \left( \frac{1}{2} + \frac{d}{2(n+1)} - \frac{1}{2} \right) j_r$$

$$= \frac{n}{n+1} \frac{dj}{2} \quad (8)$$

and since

$$\frac{(n-2)}{(n-1)} < 1 < \frac{n}{n+1} \quad (9)$$

it follows that the input to the  $B_2$  unit whose  $I$  units 'match' the input pattern will rise more rapidly than the inputs to all other  $B_2$  units and will therefore reach the trigger level first. The common feedback circuit from the  $B_2$  unit through the diodes  $K$  to all integrators acts as a holding device and prevents any other  $B_2$  unit from triggering. Thus only one of the  $M$  outputs  $O_{11} - O_{1M}$  is at the triggered level at any instant. Each output is supplied at an appropriate gain, e.g.  $O_{11} \times 1$  unit,  $O_{12} \times 2$  units, etc., to the summing amplifier  $E$  [Figures 3 and 4 (e)] to form the quantized value of the first parameter  $p_1'$ . The  $M$  quantized levels may be given any desired spacing distribution by adjusting the amplifier input resistors  $R_{11} - R_{1M}$  [Figure 4 (e)].

It has been stated that one of the important characteristics of the p.r.a.c. is that a pattern not previously encountered automatically produces the same parameter values as the pattern that it most closely resembles in all the patterns that have been presented in the past. This can easily be seen from eqn (9) which shows that the integrand produced by the two patterns with one more or one less signal than the pattern that has adjusted a set of  $I$  units is nearly the same as the input due to the latter and will therefore trigger the same  $B_2$  unit. This small pattern difference can, however, cause different  $B_2$  units to trigger when the optimum parameter values are different, so that special sets of  $I$  units are formed to resolve the pattern difference.

Having established that a certain distribution of  $a_{r+}$  and  $a_{r-}$  values is required to recognize a pattern, it remains to be

shown that the distribution can be established automatically by the pattern itself. The actual set of  $2N I$  units that is adjusted to match the presented pattern must also be selected so that the correct optimum parameter value is produced when the pattern is automatically recognized after it has been presented once during the learning process. This selection is controlled by the optimum parameter value for the process state. This value, denoted by  $p_1$  for the first parameter, is supplied to the  $F$  unit (Figure 3) which quantizes it into  $M$  levels and indicates the level by triggering one of the  $M$  outputs  $Q_1 - Q_M$ . These  $M$  levels are made to correspond with the  $E$  unit outputs and since the  $Q$  output at any instant causes a  $B_2$  unit to trigger, the value of  $p_1'$  is always a quantized representation of  $p_1$  when the switch  $S_1$  is closed during the learning process. No change takes place in  $I$  units until the performance indicator sends a negative gate signal  $G$  to the  $C$  unit when the optimum or desired performance is attained [Figures 3 and 4 (c)]. In the case of a hill-climbing parameter optimizer the  $G$  signal will indicate that the peak has been reached.

Each  $C$  unit has four binary inputs  $O_{B1}$ ,  $O_{B2}$ ,  $G$  and  $B$ , for which a '1' is represented by  $-20$  V and a '0' by zero volts approximately. The output  $O_{B1}$  from each  $B_1$  unit (Figure 3) is '1' if its input signal  $j$  is above the average  $\bar{j}$  and the output  $O_{B2}$  is '1' if its input is below the average. Each  $C$  unit has two outputs  $r$  and  $r'$  which have a resting potential of  $+20$  V that prevents the  $I$  units being driven. When  $G$  is '1' and  $j > \bar{j}$  or when  $B$  is '1' and  $j < \bar{j}$  the output  $r$  becomes  $-20$  V which allows the  $I$  units supplied by it to be driven by the drive pulses so that the appropriate  $a_{r+}$ , in the set of  $I$  units released by a  $B_2$  unit output, are allowed to increase. The second output  $r'$  becomes  $-20$  V when  $G$  is '1' and  $j < \bar{j}$  or when  $B$  is '1' and  $j > \bar{j}$  thus allowing the appropriate  $a_{r-}$  to increase.

Signal  $B$  is used to cancel the effect of  $G$  if an error is made, or it may be used in trial and error learning to eliminate all unsuccessful trial parameter values so that by a process of elimination the optimum values slowly become more probable and are finally selected.

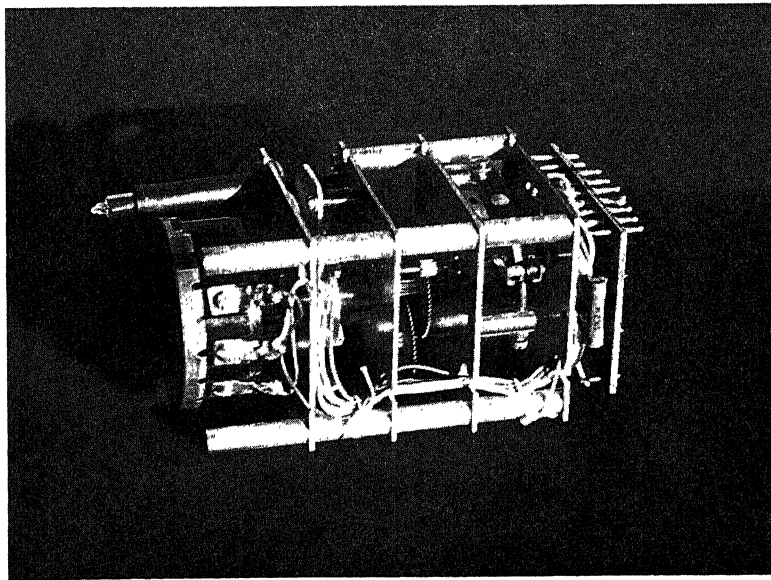


Figure 5 (a)

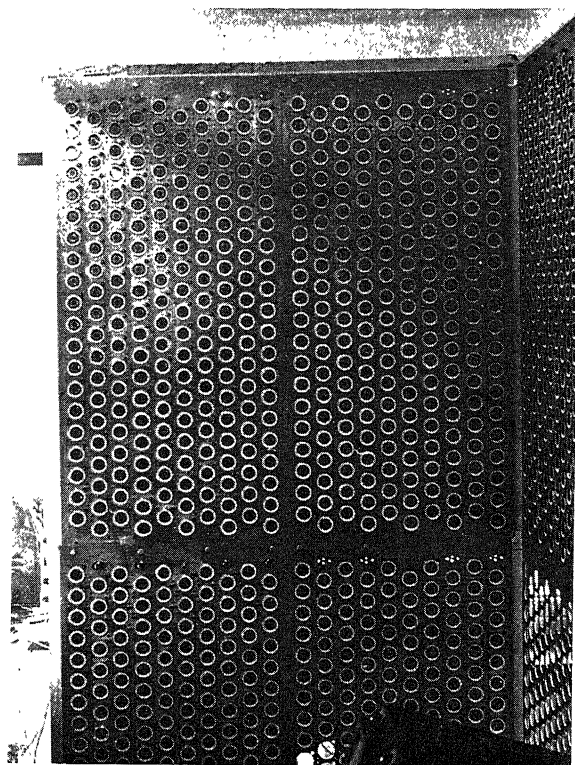


Figure 5 (b)

The magnitude of the change in gain in the sets of  $I$  units corresponding to the optimum parameter values is required to be proportional to  $1/n$ . This is achieved by driving the units for a fixed time  $T$  with pulses of a frequency that is proportional to  $1/n$ , supplied by the drive pulse generator  $L$ . A simplified equivalent circuit of the  $L$  unit is shown in Figure 4(g). In practice an electronic switch is used to generate a repetitive sequence of pulses of height proportional to  $1/n$  ( $1 \leq n \leq N$ ) and these are added to a voltage proportional to  $n$ , derived from the  $B_1$  units. The resultant waveform is used to trigger a circuit similar to the  $B_1$  unit and the output drive pulse frequency is thus proportional to  $1/n$ . The drive pulses supply all the  $I$  units,

and when both diodes in a unit are non-conducting the ratchet motors turn the potentiometer sliders [Figures 4(d), 5(a) and 5(b)] which increase the gain 'a' by  $d/2n$  in time  $\Delta T$ . The sensitivity factor  $d$  can be adjusted to match the time constants of the process by selecting appropriate values of the maximum possible drive pulse frequency, the gear ratio and  $\Delta T$ , but once  $d$  is chosen it must remain constant during the learning process.

### Termination of the Learning Process

Throughout the learning process the switches  $S_1$  in Figure 1 are closed to supply the optimum parameter values to the p.r.a.c. As the learning becomes effective the outputs of the p.r.a.c. labelled  $p_1' \dots p_k'$  become the same as  $p_1 \dots p_k$  for an increasing range of process states, but the learning process is only terminated when a criterion is reached. This criterion might, for example, be reached when ninety out of the last hundred sets of different parameter values are predicted correctly by the p.r.a.c. When the criterion is satisfied switch  $S_2$  may be changed over to connect the p.r.a.c. to the process and switch  $S_1$  opened to terminate the learning process. When the p.r.a.c. is in control it is still possible for new states of the process, requiring parameter values that have not been supplied during the learning process, to occur. The parameter values supplied by the p.r.a.c. will then be those that were the optimum values for the past process state that most closely resembles the existing state and it is improbable that they will be very far from the true optimum values.

Preliminary tests, using a small number of inputs, have shown that the p.r.a.c. behaves as predicted by the theory. Experiments with patterns consisting of 100 measurements are in preparation and the results will be reported as they become available. Several thousand  $I$  units will eventually be available for storage and Figure 5(b) shows two of the standard panels, each containing 200  $I$  units.

### References

- 1 TAYLOR, W. K. An experimental control system with continuous automatic optimization. *Automatic and Remote Control*. Vol. II. 1961. London; Butterworths

### DISCUSSION

P. H. HAMMOND, *National Physical Laboratory, Teddington, Middlesex, England*

The author's proposals are based on a steady-state open-loop learning phase, during which a given state variable pattern  $i_1 \dots i_N$  is recognized as requiring given steady-state parameter settings  $p_1 \dots p_k$  to establish optimum steady-state operating conditions. The subsequent operating phase uses the pattern-recognizing adaptive controller (p.r.a.c.) as an element in a closed feedback loop which includes the process. An immediate difficulty arises here. For a state variable pattern which happens to be present the p.r.a.c. gives as an output the best parameter setting based on experience in the learning phase. The process then responds to the changed parameter settings by producing a changing-state variable pattern. This, in turn, modifies the parameter settings and it is not at all obvious that the system will converge to a stable state. Even if it is possible to achieve stable operation in the absence of detailed knowledge of the process dynamics, the time to settle may be considerably larger than the process time constants.

The techniques discussed in the paper are certainly applicable when the inputs  $i$  to the p.r.a.c. are quite independent of the process parameter settings, i.e. process inputs and disturbances. The p.r.a.c. is then not part of a feedback loop but rather a feedforward controller.

The scheme uses a very large number of  $I$  units. It would seem that this number could be reduced by seeking efficient transformations of the state variable pattern, e.g. in terms of polynomials of which the coefficients provided inputs to the  $I$  units.

There are two small points on which I would welcome the author's comments. They are:

(1) Why is it preferable to take the second differences of the  $i$ 's to form the  $j$ 's rather than the first differences?

(2) Why is it necessary to determine the mean value of the  $j$ 's separately in unit  $A$  rather than using the sign of  $j$  itself to drive the  $B_1$  units?

W. K. TAYLOR, *in reply*

I agree with Mr. Hammond that it is not obvious that a system controlled by the p.r.a.c. will converge to a steady state, and more work is required on this problem. A steady state is not an essential condition for optimum operation, however, since the optimum parameter settings are chosen for each state. In many applications the state would, in any event, be continuously changed by unknown disturbances which may have a stronger influence than parameter changes.

It may be possible to reduce the number of  $I$  units by employing additional transformation of the state variable pattern as Mr. Hammond suggests, but only at the expense of introducing complex equipment between the process and the p.r.a.c. On the question of whether to use the first or second difference, the second difference was chosen because it gives transformed patterns that have less overlap at the  $I$  units, thereby facilitating their independent classification. If linear changes are of particular importance it may be useful to employ additional  $H$  units, taking first as well as second differences.

The difference between  $j$  and  $\bar{j}$  was taken to eliminate the effect of signal components that may be common to all  $H$  unit outputs and to suppress the random effects of low amplitude noise voltages.

A. A. FELDBAUM, *Institute of Automatics and Telemechanics, Kalanchevskaya 15a, Moscow, U.S.S.R.*

The direction of a design of such self-adjusting systems which utilized the combination of principles of automatic search and pattern recognition will, it seems to me, become of utmost importance in the near future. The very interesting work of the author is a first step in this direction and it is natural that many important questions are not considered in this paper.

The choice of the form of connection between function parameters (being extremized) and pattern values is of great importance. Instead of establishing a direct relation between extremum coordinates and regions of pattern space, it seems more convenient to pre-assign these coordinates as some functions of plant input parameters and time. The unknown coefficients in the formula for these functions can be placed in correspondence with certain regions of the pattern space.

One of the basic questions in systems with the combination of automatic search and pattern recognition is the relation between these two methods with the choice of the control action (input parameter of the plant). It is assumed in the paper that the recognizing device has first completely learned and then replaced a search system. Evidently a system as a whole will perform more efficiently if the recognizing device will participate in control before accomplishing the learning process simultaneously with the search system. Moreover, in the course of a learning process the degree of participation of the search part will decrease.

W. K. TAYLOR, *in reply*

Professor Feldbaum also asks about the possibility of forming some functions of the state variable pattern. It may be possible to obtain an advantage by introducing this complexity in some applications, but in others it is possible that the original patterns are in the best form for supplying to the p.r.a.c. The time element can be introduced into the system as it stands by supplying delayed patterns as inputs in addition to the direct inputs.

The suggestion that both the search and learning system could be allowed to operate simultaneously is quite feasible. In the extreme case the search system or hill climber can be dispensed with and a trial-and-error type of learning used with the p.r.a.c. This has been investigated on a smaller scale but has the disadvantage that the probability of selecting optimum parameters at random is initially very small unless there is a coarse quantization of parameter values to start with, the quantization becoming finer as learning proceeds.

O. L. R. JACOBS, *Department of Electrical Engineering University of Edinburgh, Edinburgh 9, Scotland*

I comment on the time scales involved in the described system.

The system in *Figure 1* of the paper is controlled in the first instance by an extremum-seeking parameter optimizer. This can only be effective if the disturbances that cause changes of state and of optimum parameter values occur slowly compared with the extremum search time. On the other hand, the p.r.a.c. is designed to recognize changes of state in a time short compared with the extremum search time. It seems then that the system to be controlled is one in which disturbances arise as sudden changes at infrequent intervals.

Could the author please give examples of systems where disturbances occur in this manner, and where there is no other way of recognizing that a disturbance has occurred than by using the p.r.a.c.?

W. K. TAYLOR, *in reply*

The disturbances may arise at relatively frequent intervals after the p.r.a.c. has taken over control, as compared with during the learning process. Examples of systems where disturbances may occur suddenly would be a process in which a raw material suddenly changes in characteristics, or a moving vehicle that suddenly enters a region of changed physical conditions. A digital computer could be used to recognize the changing states, but for complex patterns it would be many orders of magnitude slower than the p.r.a.c. which is a direct-coupled, parallel-acting network.

D. A. BELL, *A.M.F. British Research Laboratory, Reading, Berks, England*

As I am not experienced in process control I have been unable to visualize a system which has only two controllable parameters, but from the measurement viewpoint appears to have 100 degrees of freedom. Could Dr. Taylor give an example? If, as in his first example, the measurements relate to temperature distribution, surely the occurrence of local hot spots would require local remedial action rather than adjustment of the two overall parameters?

The essence of Dr. Taylor's case is that the internal relationships which reduce the apparent 100 degrees of freedom to little more than two are so complex that they cannot be explicitly stated; hence his illustrative curves in *Figure 2 (a)* have been constructed as samples of random noise. Is there not a danger that the device will store patterns which are actually random noise, and will give only a random relationship between 'pattern' and optimum values of the two parameters?

I should have thought that in many cases the nature of the pattern would be apparent—e.g. temperature gradient in a distillation column—and the specific characteristics of the pattern could be directly abstracted and used to bias the direction of search of the optimum-seeking controller. Dr. Taylor's detection of non-linearities, a process equivalent to that known as 'edging' in visual pattern recognition, is an important step, but he still assumes an apparently random distribution of edges. Is this likely in practice?

In conclusion, I would like to make it clear that I am not attacking learning machines as such, but I suggest that for industrial processes they should be avoided as far as possible. To use an animal analogy, for a comparatively narrow range of goals and environments involved in any particular industrial process it should be possible to use instinct (i.e. built-in behaviour patterns) rather than learning.

W. K. TAYLOR, *in reply*

I would like to emphasize that the p.r.a.c. is a perfectly general system that is easily adapted to any number of inputs and outputs. The example of 100 inputs was chosen to illustrate the possible complexity of a process state pattern. The system could easily be taught to recognize local hot spots and to take appropriate action when they arrive. The state patterns of a process may easily look like samples

of random noise but there are usually many physical constraints that limit the range of possible states. I agree that there are many simple cases in which the patterns of interest are fixed and known *a priori*. The controller could then be used with fixed potential dividers providing appropriate predetermined weights in place of the  $I$  units together with the  $A_2$ ,  $D$ ,  $B_2$  and  $E$  units. This combination would perform what Dr. Bell calls direct abstraction of specified characteristics and would be a pattern-recognizing controller without the adaptive elements.

Y. SAWARAGI, *Kyoto University, Kyoto, Japan*

I think that eqn (4) plays the most important role in this paper and that the controller which you stated is constructed on the basis of this equation. I would like to ask about its mathematical foundation from the viewpoint of data processing.

W. K. TAYLOR, *in reply*

I should have mentioned that eqn (4) is not the only function on which the design of the p.r.a.c. may be based. It has been found in experience to be a convenient function of the inputs but there is some reason to believe that non-linear functions of the inputs would give superior pattern discrimination. It is doubtful whether the gain would be worthwhile, however, considering the technical difficulties.

V. SOKOLOV, *Institute for Automatics and Telemechanics, Kalanchevskaya St. 15, Moscow, U.S.S.R.*

The paper presented by Dr. Taylor is very interesting as a whole. Nevertheless I would like to make some comments on that part of it which is concerned with a problem of pattern recognition. I hope that the approach which I am going to propose now is more common and would provide wider possibilities.

Let us assume that we have a set of plant patterns with corresponding points, say  $\varphi$ , in the receptor space and some single meaning transfer functions, say  $L$ , chosen from a given set of transfer functions  $U$  ( $L \in U$ ). The transfer function  $L$  will determine for every point  $f \in \varphi$  one point in a functional space. (In the most general case it will be Hilbert's space.) We say that a set of patterns  $\varphi$  is divided into  $n$  images if it is possible to divide a set of points  $\varphi$  into  $n$  subsets  $\varphi_1 \dots \varphi_n$  in such a way that the corresponding points in Hilbert's space will have the following properties.

There is only one point  $\varphi_i \in F_i$  for which the following conditions are satisfied

$$(f_i, \varphi_i) > (f_j, \varphi_i) \quad (1)$$

where  $f_i \in F_i$  and  $f_j \in F_j$

Transform (1) to the form:

$$(f_i, \varphi_i) \geq c_i \quad \text{where} \quad c_i = \min (f_i, \varphi_i) \quad (2)$$

Let the points  $\varphi_i$ , corresponding to  $(f_i, \varphi_i) \geq \max c_i$  be labelled as generalized images, and the points  $c_i$  as thresholds of identification. It is quite clear that sets of points  $\varphi_i$  and  $c_i$  will fully characterize a system of automaton images.

The points  $\varphi_i$  and  $c_i$  can be easily found by using the algorithm that was worked out at the Institute of Automatics and Telemechanics, U.S.S.R. [*Automatics and Telemechanics*, No. 5 (1963)]

The images of the automaton proposed by Dr. Taylor can be treated as well as the points  $\varphi_i$ , but the difference is that these points are determined after showing a single pattern which belongs to a particular image, and the method I am speaking about provides determination of all points in question by successive showing of patterns belonging to different images. Those patterns which are identified during the learning process (determination of points  $\varphi_i$  and  $c_i$ ) will be taken with a zero weight, and those that are not will be used to correct the corresponding coordinates of points  $\varphi_i$  and  $c_i$ .

Therefore the last method provides a higher degree of accuracy, and in addition it can be used in the case when  $f_i$  is expressed in any desirable form.

W. K. TAYLOR, *in reply*

I agree that it is possible to obtain greater accuracy in pattern recognition by choosing more complex functions of the state variables that in effect discard less information about the process state than do the functions chosen in the paper. The question that should be asked, however, is whether the resultant increased cost and complexity of the controller is justified by the improvement in control. The answer will depend upon the particular application and may only be obtained after detailed experimental investigation.

In reply to Mr. Sokolov's remarks on the training procedure, I would like to emphasize that it is, in general, necessary to show patterns belonging to different 'images' and the two methods would appear to be equivalent. Thus, although a large set of related patterns is given the same classification by presenting a single member, it is possible to subdivide the set by successive presentations of members that are required to produce different outputs.



# STELLA: A Scheme for a Learning Machine\*

J. H. ANDREAE

## Summary

A scheme for a learning machine is described. In a basic exploratory mode the machine searches its environment. Learning enables it to profit from those action sequences which lead to reward. By correlating the changes in its environment with its actions, the machine can extract invariant features and use them to guess profitable actions when its learned sequences fail. An internal mode of operation is described in which the machine explores the possibilities of its future actions with a view to modifying its performance. The learning automaton, STELLA, is being constructed in the form of a mechanical tortoise which takes its name from its laboratory origin.

## Sommaire

On décrit le schéma d'une machine à apprentissage. Dans un mode fondamental d'exploration la machine cherche son amb'ance. L'apprentissage la rend capable de profiter de celles des séquences d'action qui conduisent à la récompense. En corrélant les variations de l'ambiance avec ses actions, la machine peut extraire les traits invariants et les utiliser pour deviner les actions profitables quand ses séquences apprises lui font défaut. Un mode interne de fonctionnement est décrit dans lequel la machine explore les possibilités de ses actions futures en vue de modifier ses performances. L'automate à apprentissage, STELLA, est en cours de construction sous la forme d'une tortue mécanique qui tient son nom du laboratoire d'origine: Standard Telecommunication Laboratories Ltd.

## Zusammenfassung

Der Aufsatz behandelt die systematische Anordnung eines lernenden Automaten. Durch einen grundlegenden Suchvorgang „erforscht“ die Maschine ihre Umgebung. Lernen befähigt sie, diejenigen unter den möglichen Handlungsfolgen (Schritten) auszusuchen, welche zu einer „Belohnung“ führen. Durch Korrelation der Veränderungen in ihrer Umgebung mit ihren Schritten kann die Maschine nicht veränderliche Merkmale feststellen und gewinnbringende Schritte „erraten“, wenn keine der gelernten Schritte anwendbar sind. Ein interner Arbeitsvorgang wird beschrieben, mit welchem die Maschine die Möglichkeiten von zukünftigen Schritten im Hinblick auf Abwandlungen ihrer Arbeitsweise erforscht. Die Konstruktion des lernenden Automaten hat die Form einer mechanischen Schildkröte; der Name STELLA kommt von „Standard Telecommunication Laboratories“, wo er gebaut wurde.

## Introduction

STELLA is a scheme for a general-purpose learning machine. For convenience, the scheme is being tested in the form of a mechanical 'tortoise' moving about on a level floor which can observe the approximate distances and angles of obstructing walls. The scheme is intended to be unspecific to the particular type of environment in which the machine operates, and other machines and environments are under consideration.

\* Progress made with the STELLA learning machine since this account was submitted on 23<sup>rd</sup> August 1962 is the subject of a paper being prepared for publication.

There would appear to be two main approaches to the design of learning machines. In the one, a rigid control system performing a strictly determined task is modified to make it less rigid, more flexible and adaptive. In the other, a general-purpose learning machine is designed and is then modified to make it more specific to particular problems. STELLA falls within this latter category and the object of experimenting with the STELLA scheme is to explore ways in which it can evolve into a more specialized system. The importance of this process of 'progressive segregation' of parts is emphasized by Bertalanffy<sup>1</sup>.

STELLA, the tortoise, is a self-propelling and self-steering trolley which can move about on a level floor. Walls of standard height having illuminated upper rims obstruct its movements in various directions. Information about the angular positions and distances of these walls is obtained by photocells mounted around the circumference of the trolley. There are eight of these photocells on the machine which is being constructed. *Figure 1* assumes six only.

## The Exploratory Mode

Unless its memories are 'primed' by the experimenter, STELLA starts with no information about its environment. In these circumstances, and whenever the higher level of operation fails to decide its actions, STELLA proceeds in an exploratory or search mode. This mode is governed by a random generator in such a way that the machine tends to perform different actions when receiving the same information repeatedly from the photocells of its eye.

Each photocell of the eye has a fixed threshold which the light input to the photocell must exceed for the information from that photocell to change from  $-1$  to  $+1$ . Thus each photocell contributes one 'bit' to the input binary pattern from the eye—the eye pattern.

The actions of STELLA are restricted, in *Figure 1* and in the constructed machine, to four movements: forward, forward to the left, forward to the right, and reverse.

The short-term memory shown in *Figure 1* consists of a  $6 \times 4$  matrix of binary elements represented by the intersections of the six column lines from the eye and the four row lines to the dissimilarity detector (DD). Each row line is associated with one of the output movements, forward, left, right or reverse.

When the eye 'sees' a pattern, the DD selects that row in the short-term memory along which is stored the pattern least similar to the eye pattern. The motor output control then performs the output movement or action corresponding to the selected memory row. The continuous random improver of the memory switches, by random choice, some of those elements in the selected row which are not the same as the corresponding elements of the eye pattern, until, due either to the movement causing a change in the environment or to the random improvement of the row, that row ceases to be the least similar to the

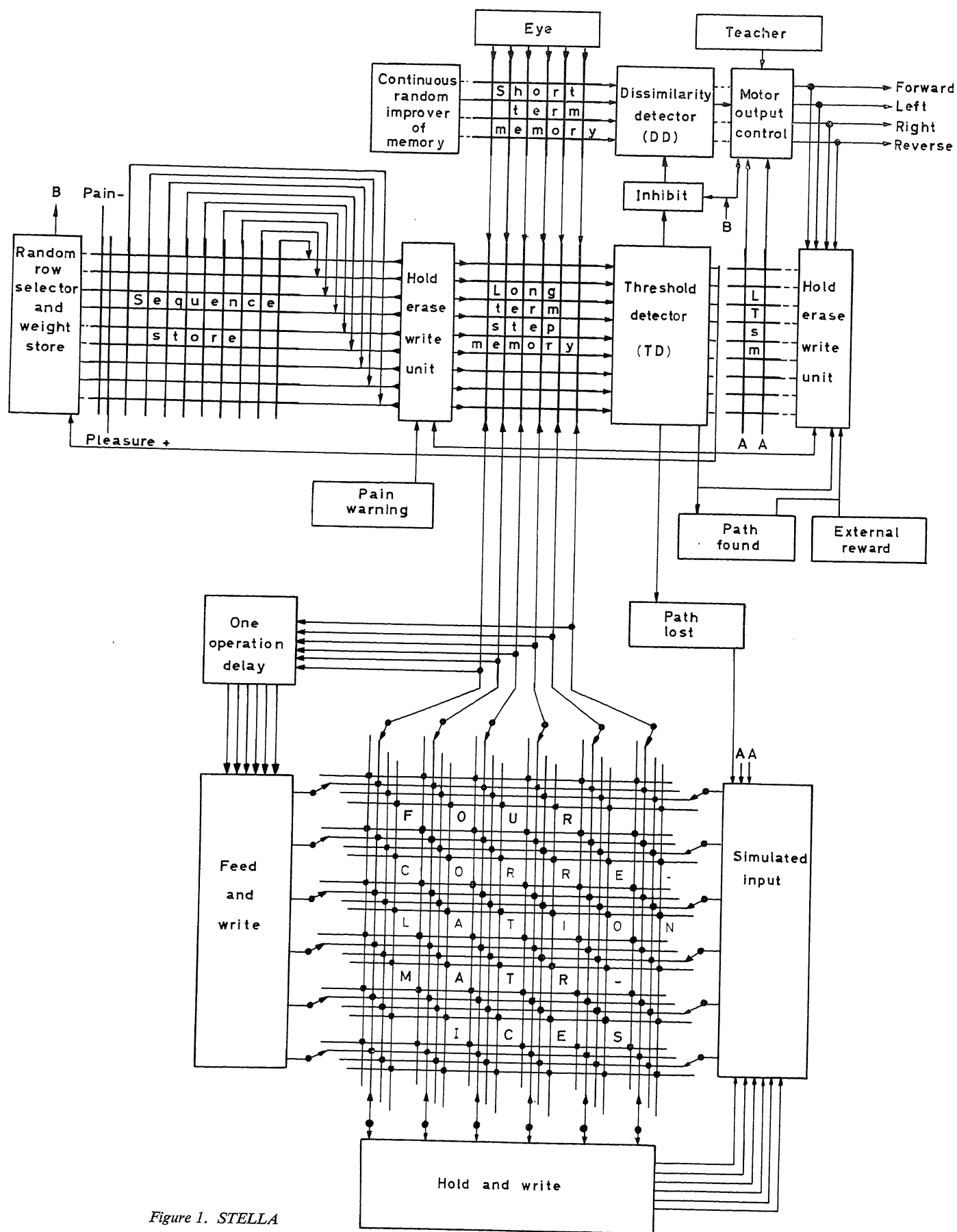


Figure 1. STELLA

eye pattern and the *DD* selects a different row. The selection of least similar rows follows the procedures of Steinbuch<sup>2</sup>; the random improvement is superimposed to provide a short-term memory which eliminates repetitive reactions to the eye pattern: STELLA should not push persistently against walls and should break out of repetitive cycles.

### Learning by Reward

Patterns and actions can be stored in the two parts of the long-term step memory shown in *Figure 1*. At the same time, and in the same row, a connection can be made in the sequence store to indicate that the pattern and action of that row was followed by a pattern and action stored in another row, or by reward. Each intersection between a row line and column line in the sequence store represents the following of the action of that row by the pattern in the row connected to that column line; or, if it is the reward column (marked 'pleasure' in *Figure 1*), the following of the action of that row by reward. There may be different kinds of reward with separate columns, and there may be a 'pain' column; these are considered in later sections.

When an eye pattern is received, it is compared with the patterns stored in the long-term step memory by the threshold detector (*TD*). The *TD* associates any row of the memory which has a pattern more similar to the eye pattern than the current threshold of similarity demands. Consider three cases:

(a) The *TD* associates no row, but the action prescribed by the *DD* leads to reward. The eye pattern and the subsequent action are stored in a vacant row of the long-term step memory and a connection is made in the sequence store between that row and the reward column. The row is given unit weight in the weight store.

(b) The *TD* associates one row. The *TD* overrides the *DD* and causes the motor output control to perform the action stored in the associated row of the long-term step memory. At the same time, suppose that there is a connection in the sequence store which indicates that on a previous occasion the action was followed by the pattern in another row of the long-term step memory. The weight of this other row is increased in anticipation of its being seen again. Also, while the action is being performed, the preceding pattern and action are held tentatively in a vacant row of the long-term step memory; in fact, every pattern and action is held tentatively for one cycle. Now, the action performed, several situations may ensue:

(i) If there is no association of the new eye pattern by the *TD* and no reward, the weight of the anticipated row is reduced to what it was, the tentatively held pattern and action are disregarded, the weight of the last row to be associated by the *TD* is reduced, and the machine STELLA reverts to the search mode.

(ii) If reward is received, the weight of the associated row is increased, a connection in the sequence store is made to the reward column, and the tentatively held pattern and action are stored with unit weight in a row with a sequence connection to indicate the row which followed.

(iii) If the anticipated row is associated by the *TD*, its weight is increased further in confirmation, the tentatively held pattern and action are held for another cycle and the next step is entered.

(iv) If an unanticipated row is associated by the *TD*, the weight of the anticipated row is decreased and the weight of the associated row is increased.

(v) If the *TD* associates more than one row, one has, as for (c).

(c) The *TD* associates more than one row. Each of the associated rows has a weight depending upon how many times it has been used in the past and each row is connected to the reward column through a series of connections in the sequence store. Each connection in this 'circuit' of the sequence store between the row and the reward column represents a possible step. If each connection is imagined to have an electrical resistance and the column and row lines of the sequence store are imagined to be wires, then the total resistance between an associated row and the reward column is a measure of the number of steps expected to lead to reward. The random row selector chooses one of the rows associated by the *TD* on a chance basis biased towards high row weights and short paths to reward. Once a single row has been chosen, the situation is the same as for (b).

*Figure 2* is a flow diagram of the processes outlined above, together with other processes which will be described.

The threshold of the *TD* determines the permitted latitude in the association of row patterns with the eye pattern. If the threshold of similarity is low, the behaviour will tend to be illogical and opportunity will exist for the discovery of shorter paths to reward. When STELLA is being particularly successful in obtaining rewards, it is necessary for the threshold to be kept high so that previously established paths will be followed more precisely the more successful they prove to be. However, it is not desirable that a pattern should be completely disregarded in favour of search behaviour because the threshold happens to be just too high. Therefore, it is proposed to have a slow decline in the value of the threshold with time, to cause the threshold to be restored to a maximum when reward is received, and to allow the *TD* to depress its own threshold (temporarily) by not more than a predetermined amount below its current value, if no pattern is associated.

'Forgetting' is necessary to remove those steps which are less used and, therefore, less useful so as to make room for more important steps. To do this the weights of the rows in the long-term step memory are given slow time decays. When a weight falls below some prescribed value, the contents of the row are erased and all connections to and from the row in the sequence store are broken.

The 'teacher', shown in *Figure 1*, is a control which overrides the motor output control to enable the experimenter to speed up the learning process by forcing STELLA along profitable paths.

### Pain

There will be some values of the parameters controlled by a learning system which must be avoided. For example, in the control of a chemical plant it may be known that certain mixtures of reactants are explosive and that these must be avoided; again, in a traffic control system it is likely that there will be a minimum safe distance of approach of vehicles. In the tortoise floor-wall system one can imagine holes in the floor which the tortoise must avoid in order to survive. Clearly it is inadequate to provide learning by reward only in a machine which can act by trial and error in its environment if forbidden conditions are to be avoided. It must be taught or be programmed to avoid these conditions and the information which determines this avoidance will be called pain. Responses to pain cannot be maintained by experience in the way responses to reward are reinforced.

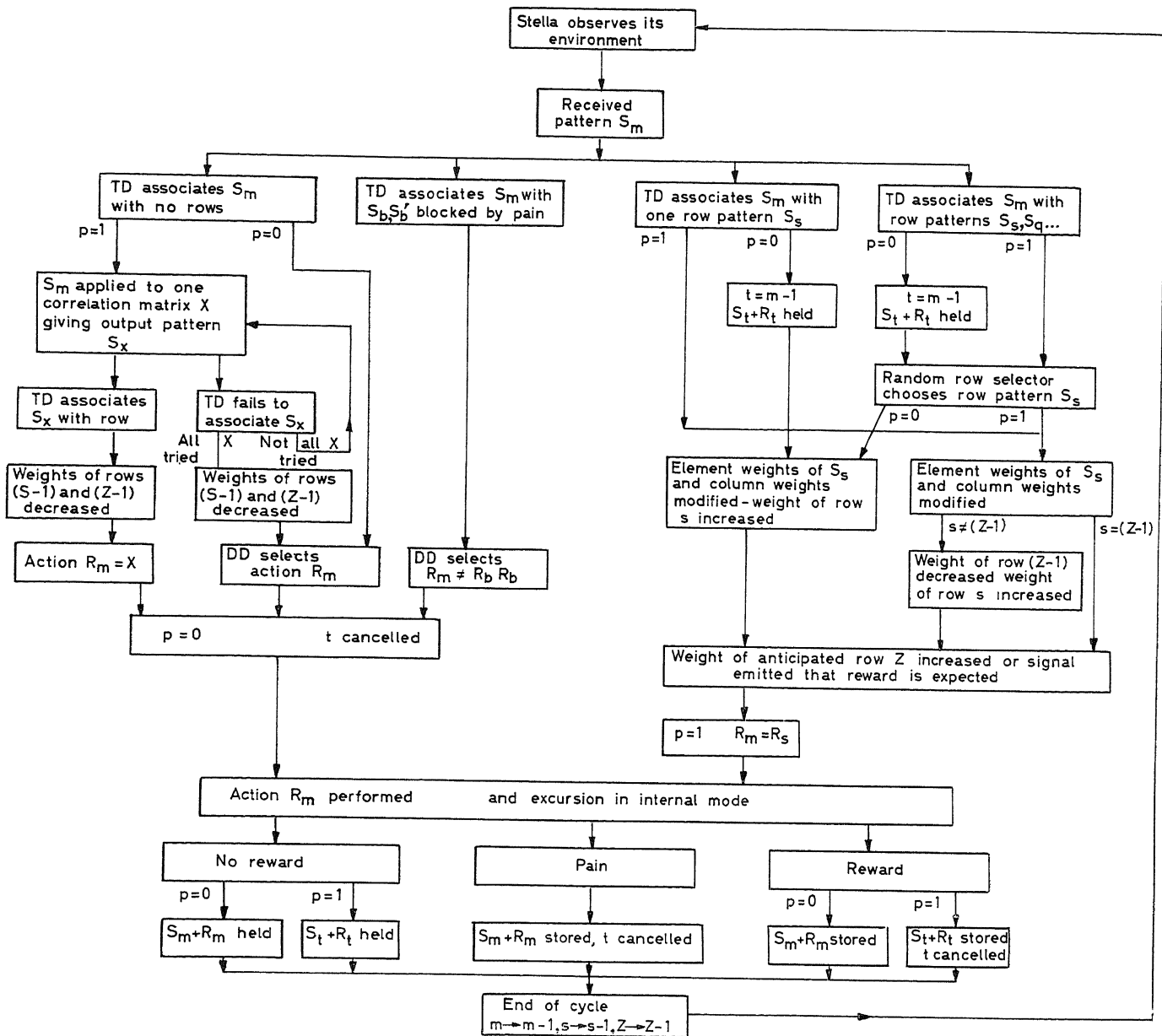


Figure 2

A traffic control system could not be allowed to maintain a safe distance of approach between vehicles by regular experiments with unsafe distances.

Pain is introduced into the STELLA scheme of Figure 1 by the addition of a 'pain' column to the sequence store and by an externally controlled 'pain warning'. In order to understand the implication of the negative sign attached to the pain column, which may be compared with the positive sign ascribed to the pleasure, or reward column, it is convenient to think in terms of the actual electrical method employed for the operation of the random row selector. In order to bias the decision of this selector according to the length of the path (number of steps) to reward, the connections in the sequence store are resistors and the row line in question is earthed, while the reward column line

is connected to a positive battery voltage. In this way a current flows from battery positive to the earthed row which is inversely proportional in magnitude to the resistance of the path, that is, the length of the path. Now some connections to the pain column will be preprogrammed into the machine and some will be formed as a result of the pain warning. If the pain column is connected to a negative battery voltage, the currents to this column, in so far as they pass through the same resistors as the current from the positive reward column, will oppose the currents from the pleasure column and decrease the probability that the random row selector will choose the row in question. Actions expected to lead to pain will be inhibited according to how many steps are expected before pain may be encountered. If the pain currents exceed the positive currents, it is arranged for the action pre-

scribed by the row to be blocked. That action is forbidden for that step.

The connections to the pain column are made in the same way as for those to the reward column, but allowance is made for preprogrammed connections to the pain column to be established for some rows with permanent (non-decaying) weight.

### Specialization

In order that STELLA should learn paths out of painful situations, the cessation of the pain warning should be treated as a reward and an additional positive reward column is required for the sequence store. This column would become connected to the positive battery voltage when a pain warning was received and it is logical to arrange for the other reward column to be disconnected simultaneously. Suppose that the machine is given a goal and the pleasure reward indicates the achievement of this goal. If the implication of a pain warning is sufficiently serious, as it is intended to be, then the final attainment of the goal is subject to a temporary diversion to the subsidiary goal of escaping from the painful situation. Similarly, if the system has to ensure the supply of its own power or the supply of raw materials to the process it is controlling, it may be essential for this secondary goal to be attended to at the expense of progress towards the prime goal. This logical interdependence of the operation of the reward columns forces upon us the specialization of the system to match its environment. The machine must be told, to begin with, what is required of it. It must also be given some rules to prevent it from destroying itself or others. Natural selection is too costly.

What other kinds of specialization have to be introduced? The scheme of *Figure 1* was designed to be unspecific to its environment. The connections from the eye can be interchanged so that different photocells control different columns of the memories and after a time the system should adapt itself to the new arrangement. The same applies to the connections to the output actions. But this interchangeability of connections presumes that each photocell is providing equivalent information; otherwise patterns stored in the long-term step memory will give undue importance to digits on some of the columns at the expense of those on others.

It seems possible to arrange for STELLA to compensate for the unequal importance of digits in the eye pattern. Each column is given a variable weight which determines the bias applied to the appropriate element of a pattern associated by the *DD* or *TD*. The column weight is decreased every time the *TD* disregards the corresponding digit of the eye pattern in associating a pattern in the long-term step memory with the eye pattern. After a time some columns will be contributing little to the operation of the machine and the experimenter can change the position or other property of the respective photocells in order to achieve an arrangement in which these photocells contribute their full share. The machine could, of course, be programmed to make such changes itself. However, the experimenter might find that more radical changes were needed. For example, it might be more effective to start with 100 photocells logically connected to give only the six outputs, these outputs representing more specific characteristics of the environment. See, for example, Lettvin *et al.*<sup>3</sup>

The variation in the column weights should represent the relative importance of the information arriving on the various

columns, when the importance is averaged over a large number of steps, and it should indicate the efficiency of the photocells as receptors of information. The relative importance of elements of a pattern stored in the long-term step memory will depend, not only on the efficiency of the receptors, but also on the situations in which the pattern is used. Suppose that the experimenter decides to reward STELLA (the tortoise) every time it reverses on approaching a wall. As things have been described, STELLA would have to store a number of patterns representing the various situations in which it approaches a wall, each of these patterns being coupled with the action 'reverse'. If the machine is not being very successful in obtaining reward and the threshold is low, then it may be lucky enough to use one of these patterns in a situation which does not quite correspond to the remembered situation, but this is not possible when it is more successful. Now let each element of each pattern in the long-term step memory have variable weight, the weights for each pattern being normalized so that the *TD* will still just associate this pattern with the eye pattern if it is identical and if the threshold is at a maximum. The element weights are modified each time the row is selected by the random row selector, disregarded elements having their weights lowered. The experiment envisaged above will lead to the establishment of a pattern in the long-term step memory coupled to the action 'reverse' with its element weights adjusted so that only those digits of the pattern, which indicate the presence of a wall ahead, contribute to the association procedure of the *TD*. A crude kind of generalization takes place so that a number of possible patterns representing a particular situation are accepted and remembered as a single pattern.

### The Correlation Level

It was stated above that to begin with the machine must be told what is required of it, but this is not strictly true. It can explore its surroundings and learn something about their characteristics while the experimenter is still making up his mind about what he wants the machine to do. STELLA does this constantly by means of its correlation matrices, which are shown, drawn one upon the other, in *Figure 1*. Each of the four matrices is associated with one of the actions.

Every eye pattern received is applied to the columns of the matrix appropriate to the action which has just been performed. The same pattern is then held for one step (one action) and applied to the rows of the matrix corresponding to that action. Thus, for any action, the appropriate matrix has the initial pattern applied to its rows and the resultant pattern applied to its columns. So long as the action has not been prevented by the external environment, each element of the matrix has its 'value' increased (decreased) if its row and column have the same (opposite) binary values (+ 1).

When STELLA has explored the environment for some time, the correlation matrices should contain experience of the way in which its actions transform the observed environment. Invariants in these transformations will be reinforced by persistent increases or decreases in the values of particular elements of the matrices. For example, in the tortoise floor-wall arrangement one would expect the matrix corresponding to the forward movement to contain the information that an observed pattern will move past the tortoise from front to back.

The correlation matrices are used if the *TD* fails to associate a pattern when a remembered path is being followed. The

sequence of operations is shown in *Figure 2*. The unassociated pattern is applied to each of the correlation matrices in turn until the *TD* associates one of the transformed patterns. If the *TD* associates one of the transformed patterns, then the action corresponding to the matrix which effected the transformation is performed, as a hopeful guess based on experience. The transformation by one of the matrices is carried out by matrix multiplication of the pattern binary vector by the non-binary matrix, the signs of the components of the product vector determining the binary elements of the transformed pattern.

If the guesses resulting from the correlation matrices sometimes enable STELLA to rediscover the paths which it loses, they will contribute to the efficiency and speed of learning. There is, however, a more significant way in which the correlation matrices can take part in predictive forecasts. This is the subject of the next section.

### The Internal Mode

If the reception of eye patterns and the performance of actions are blocked, an internal mode of operation can be envisaged which might be called 'dreaming'. The last eye pattern to get through before the blockage leads by the *DD* or the *TD* to the selection of an action but, instead of the action being performed, the corresponding correlation matrix is used to

transform the eye pattern into a 'guessed' second pattern. This second pattern is now treated as a new eye pattern, the *DD* and *TD* select a correlation matrix to form a third pattern, and so on. The 'dreams' may lead sometimes through the random excursions of the *DD* and sometimes more logically by the *TD* through the remembered paths of the long-term step memory to occasions of pain and pleasure.

In the internal mode of operation, STELLA is exploring the possibilities of future actions by using the information stored in the correlation matrices and in the long-term step memory to anticipate the effects of its postulated actions. If time is allotted during the performance of each action (see *Figure 2*) for short excursions in the internal mode, and if some variation of the weights of rows in the long-term step memory is permitted when the excursions anticipate reward or pain, then STELLA can modify its own actions according to the machine's predictions.

### References

- <sup>1</sup> VON BERTALANFFY, L. An outline of general systems theory. *Brit. J. Phil. Sci.* 1, 2 (1950) 134
- <sup>2</sup> STEINBUCH, K. Die Lernmatrix. *Kybernetik* 1, 1 (1961) 36
- <sup>3</sup> LETTVIN, J. Y., MATURANA, H. R., MCCULLOCH, W. S. and PITTS, W. H. What the frog's eye tells the frog's brain. *Proc. Inst. Radio Engrs, N. Y.* 47, 11 (1959) 1940

# Automatic Control Learning Systems (in the Light of Experiments on Teaching the Systems Pattern Recognition)

M. A. AIZERMAN

## Summary

The paper is devoted to the problem of teaching automatic systems pattern recognition. At the basis of the experiments that were conducted with regard to teaching, lies a profound and original hypothesis relating to the teaching process—a so-called '*compactness hypothesis*'. This hypothesis is interesting in that it opens a path to the mathematical description of the essence involved in the teaching process, which description is not attached to any concrete technical realization of the learning system nor to any concrete pattern type. At the same time, it prompted the experiments described in the present paper, and also made it possible to comprehend already-known studies on teaching.

The fact that the experiments on teaching described in this paper were conducted on a universal (multi-purpose) computer is naturally tied in with the presence of such an hypothesis. The '*compactness hypothesis*' makes it possible to formulate various teaching algorithms. Two of them (the algorithm of random planes and the algorithm of *potential* surfaces) were verified experimentally and yielded good results.

## Sommaire

Ce rapport traite le problème de l'enseignement de la reconnaissance automatique de la structure des systèmes. Une hypothèse profonde et originale, concernant le processus de l'enseignement, et dite 'hypothèse de compacité' a été émise et a servi de base d'expérimentations. Cette hypothèse ouvre la voie à une description mathématique de l'essence même du processus d'enseignement, indépendamment de toute réalisation concrète de système d'apprentissage et du type de structure concrète à reconnaître. Elle a permis les expériences décrites et a permis de mieux comprendre les études connues sur la même question.

Les expériences décrites dans ce rapport ont été effectuées sur un calculateur universel. L'hypothèse de compacité a rendu possible la formulation d'algorithmes d'enseignement. Deux d'entre eux (l'algorithme de plans aléatoires, et l'algorithme de surfaces potentielles) étaient vérifiés expérimentalement et avaient donné de bons résultats.

## Zusammenfassung

Der Aufsatz befaßt sich mit dem Problem, automatischen Systemen die Zeichenerkennung zu lehren. Die Versuche darüber bauen auf einer tiefeschürfenden neuartigen Hypothese des Lernvorganges auf, der sogenannten „compactness hypothesis“ (umfassende Hypothese). Diese Hypothese ist deshalb interessant, weil sich durch sie das Wesen des Lernvorganges mathematisch beschreiben läßt; diese Beschreibung ist weder an eine bestimmte technische Verwirklichung des Lernsystems noch an eine bestimmte Zeichentypen gebunden. Sie war auch der Ausgangspunkt für die hier beschriebenen Versuche und ermöglichte die Einbeziehung bereits bekannter Untersuchungen über das Lernen.

Die Tatsache, daß die hier vorgelegten Lernversuche auf einer Universal-(Mehrzweck-)Maschine durchgeführt wurden, hängt natürlich auch mit dieser Hypothese zusammen. Die „compactness

hypothesis“ ermöglicht es, verschiedene Lernalgorithmen zu formulieren, von denen zwei, der Algorithmus der zufälligen Ebenen und der der Potentialfelder, mit guten Ergebnissen geprüft wurden.

## Introduction and Setting up of the Problem

The term 'learning automaton' has varying interpretations, depending on whether the existing automatic control systems or automatic control systems of the near future are in mind.

In learning control systems, which are beginning to be employed more and more widely in present-day techniques, by teaching, one generally means the process involving the gradual improvement of the system's operation under definite operational conditions, or the capacity of the system for varying the adjustment changing the operational conditions. Such instruction is accomplished through the action of 'reward' and 'penalty' signals, which evaluate the results of the control system and are employed in the system for the purpose of seeking optimum alignment. These signals are sent out by an operator or by another automaton, but in any case the goal of the instruction is the alignment of dynamic system characteristics.

The successful solution, by present-day techniques, of problems of this type, lays the basis for expecting, in the near future, a considerably more intensive and delicate utilization of the teaching process in automatic control systems.

The possibilities involved in teaching living beings, even including those that stand on the lowest rungs of the evolutionary ladder (for example, worms), are still incommensurably greater than the possibilities of teaching the most advanced automatic control systems. How far is it possible to advance in imitation of the learning process of the living brain by means of the automatic machine? In the near future, may a material expansion be expected in the possibilities of teaching automatic control systems? Many expressions of opinion have been published, in recent years, containing general points of view on questions of this nature, but, unfortunately, throughout the entire world, very few experiments have been conducted and published, that would be in a position to give substantiated answers to such questions and offer a glimpse into future techniques that could be used in teaching automatic control systems.

Even though, in the setting up of such experiments, the very broadest understanding of the term 'teaching' is kept in mind (teaching accommodation to changing conditions, teaching the most favourable behaviour, and so forth), the experiments are usually conducted under conditions of teaching the automatic



machines the recognition of patterns. The capability of the living brain for learning pattern recognition is one of its most delicate capacities, and any advance in imitation, by the automatic machine, of this feature of the living brain's activity immediately indicates the paths for the reproduction, by the automatic machines, of other, simpler learning processes, too.

The present report is devoted to a consideration of this study, which was carried out in the laboratory (under the supervision of the author) of the Institute of Automation and Telemechanics, by Braverman<sup>1</sup>, with the participation of O. A. Bashkurov and I. B. Muchnik, and is tied in with the series of problems solved in the papers of Rozenblatt<sup>2</sup>, Selfridge<sup>3</sup>, Steinbuch<sup>4</sup>, Bongard<sup>5</sup> and others. The studies with which this paper is concerned differ from those of References 2, 3 and 4, with regard to the ideas on which the teaching is based, and also, because Rozenblatt, Selfridge and Steinbuch constructed, for the purpose of imitating the teaching process, specialized automatic machines (Rozenblatt's Perceptron, Selfridge's 'Pandemonium', Steinbuch's learning matrix, etc.). On the other hand, this paper's experiments were conducted on universal electronic machines.

When speaking of teaching the recognition of visible patterns, the following two differing processes are in mind:

(1) The instructor knows more than the student. The teacher is not only capable of recognizing the patterns, but has also reasoned them out in advance, as must be done to conform with each concrete, partial case. Knowing in advance which visible patterns the pupil will have to recognize, the instructor explains to the pupil how it must be done, i.e., he assigns a programme, so to speak—the recognition algorithm. The pupil memorizes this programme and carries out the recognition of the visible patterns afterwards shown to him in conformity with it. Such understanding of the term 'teaching' does not indicate any interest for the purposes of automatic control. This understanding brings one only to the usual programming problems which always arise in the theory of digital computers.

(2) There is yet another teaching process, another capability of the living brain, that is not understood up to the present time. If a pupil, not knowing the alphabet, is shown the various outlines of several alternatives of the letters 'a' and 'b', and is told that this letter is 'a' and that letter 'b' then, after a certain period of time, the pupil will be able to distinguish the letters 'a' from the letters 'b' and, at the same time, distinguish not only those that were shown to him previously, but also all the remaining outlines of these letters.

An automatic imitation of precisely such a teaching process constitutes our purpose, if teaching in control systems is borne in mind.

Up to this point, in speaking of a 'visible pattern', the author has depended merely on the intuition of his audience. The concept is now established more accurately.

By a visible pattern, is meant an infinite set of visible perceptions; for example, plane figures, representations, which is remarkable with regard to the following: people, seeing only some elements of this set, confidently attribute to it other elements not previously seen.

Thus, for example, a set of portraits of all the persons in attendance here, does not constitute a pattern because if a person not present here were shown some of these portraits, he would not be able, afterwards, on seeing a new portrait, to tell us whether the face represented on it was that of a participant at our gathering. The concept of a 'male portrait', however, does

constitute a pattern, because, even though no one has seen all the male portraits, each one of us, upon seeing a new portrait, would be able to say, with a high degree of assurance, whether it represents a male portrait. Other examples of patterns are represented by the concepts of the 'number two', the 'letter a', a 'circle', a 'landscape', etc.

When in teaching a machine the recognition of visible patterns is spoken of, the following experiment is in mind. Imagine that there is a photo-field made up of a large number of photo-elements. Images are projected on to the photo-field—the elements of some kind of patterns, for example, some sort of outlines of the numbers two and three. In each projection, from the output of each photo-cell, there is taken off a current of definite intensity, i.e., the entire photo-field reaches a certain state. From each photo-cell, wires are led up to the device that we call an automaton.

In the teaching process, from the entire uncountable (non-enumerable) amount of pattern elements, for example, of the outlines of numbers 2 and 3, a few, say 20, are selected and shown to the machine by projecting them on to the photo-field. In each display, it is communicated that by pressing a definite button, a two, a three, and so forth have been demonstrated. After all the selected numbers have been shown, the teaching process is concluded, and the pushbutton keys are switched off (cancelled).

Following this, on the photo-field there are projected various other outlines of twos and threes that had not been used previously. The machine, by some sort of conventional signal, for example by switching on an appropriate panel indicator, gives an answer to the question: 'What is this?' One will be able to say that a teaching process has been produced if the machine, in an overwhelming majority of cases, gives the correct replies. The requirement is that, without changing anything in the machine circuit or in its programme, the same test with new patterns can be repeated. For example, it should be possible, without changing anything in its programme, to take a machine that has been taught the recognition of numbers, and by means of the same process, reteach it the recognition of letters or geometric figures, or teach it to distinguish male and female portraits, and so forth.

### Compactness Hypothesis

By the very nature of the problem, it is not known in advance which concrete patterns the automatic machine will have to learn to recognize. For this reason, a set of signs cannot be included in the programme that is suitable for the recognition of a given concrete pattern. Only two paths which could lead to the solution of the problem are seen:

(1) The machine programme would contain so many of the most varied signs, which have either been fed into it previously or developed by the machine itself, that this set of signs would be sufficient for the recognition of various patterns.

(2) There should be found and fed into the programme some kind of single, or small number of universal, 'signs', that lie at the basis of the very concept of 'pattern', and which are not specific for the recognition of concrete patterns.

It would seem that such a universal sign exists, that it is closely associated with the above-mentioned concept of 'pattern', and that it may be formulated in the form of a hypothesis, which was called a 'compactness hypothesis'.



In *Figure 1*, is the representation of several patterns for the outlines of numbers five and three. If these pattern elements were shown to a large number of people, one by one, and they were asked: 'What number is this?', then the decision would be practically unanimous: in the upper row, fives are represented and in the lower one, threes. With reference to the numbers represented in the centre row, it is certain that there would be no unanimity. In going from one 'five' to another, all the intermediate pattern elements will also be perceived by the majority of viewers as fives, and in a smooth transition from a five to a three, there will necessarily appear pattern elements with regard to which opinions will be divided, and it will no longer be possible to assemble an overwhelming number of voiced opinions towards the clarification of what number is being represented. In addition to this, if the representation of the figure five (or three), is altered to a negligible degree in any direction, then the new pattern element will likewise be perceived as the number five (or three).

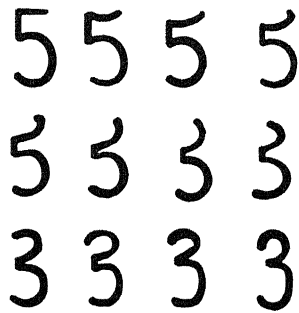


Figure 1

The compactness hypothesis assumes that all the patterns, in fact, possess these properties; namely (a) that a smooth transition is always possible from one element of a given pattern to another; this takes place in such a way that all the intermediate pattern elements will be perceived as elements of this pattern. On the other hand, it is not possible to have a smooth transition from the elements of one pattern to the elements of another without the development, 'on the way', of pattern elements, regarding which opinions would be divided, and, without having a lack of unanimity with respect to which one—the first or the second pattern—to relate them, and (b) the pattern limits are not exceeded in the case of a minor pattern element alteration in any direction.

Let an  $n$ -dimensional space be matched up with a photo-field containing 'n' photo-cells in such a way that along each coordinate axis there is laid off the condition of one of the photo-cells. In that case, a point in this space corresponds to each photo-field condition, and will be called a receptor space.

Now project some type of pattern element on to the photo-field. All the photo-cells will assume some condition, that is, a definite photo-field state will correspond to this pattern element, which means that a single point in the receptor space will be obtained.

The compactness hypothesis confirms that a domain in the receptor space corresponds to all the pattern elements that comprise some type of pattern, and that different domains correspond to different pattern elements, without having any common points.

Keeping this  $n$ -dimensional pattern in mind, it will be

represented conditionally on a two-dimensional plane. In that case, it can be visualized that two separate domains exist objectively in that plane—one of them corresponds to the first pattern, for example to the figure 'five', and the second one to the second pattern, for example to the figure 'three'. The domain boundaries are not known in advance.

In the teaching process, a certain amount of random-selected elements of the first and second pattern is shown, i.e., a certain amount of random points, respectively, from the first and second domains. The problem of the automatic machine consists in that, knowing only these points, it should be able to lay out the surface of separation in such a way that the domains appear distributed along various sides of this surface.

When, at the end of the teaching process, there will appear new pattern elements or, what amounts to the same thing, new points in the receptor space, the machine will be able to answer the question as to what patterns these points belong to, dependent only on what side of the surface of separation these points lie.

If the surface of separation is successfully drawn, and the domains are completely divided by it, then the machine's replies will always be correct, no matter how many new pattern elements belonging to these patterns were shown.

The aim, at the present time, consists in finding automatic algorithms for laying out the surface of separation, proceeding only from a small number of random selected points, from domains whose boundaries are unknown.

### Teaching Process Algorithms

Two automatic algorithms were suggested for the purpose of solving the problem regarding the construction of a surface of separation for emerging random points, in the case where the boundaries which underlie the domain division, are unknown.

*First algorithm—algorithm of 'random planes'.* Assume that at first, in the teaching process, two points appeared, and that the machine was informed that they corresponded to different patterns (for example, the figures three and five were shown). In that case, the machine traces a plane that is selected at random, with a single, limiting condition that it must separate the emergent points [Figure 2 (a)]. If, subsequently, a point appears once more, and it turns out that two points, which belong to varying patterns, are not divided by a plane, then a new, random plane is drawn that divides them [Figure 2 (b)], and so on, during the entire teaching process. In this manner, at the end of the teaching process, a large number of traced random planes [Figure 2 (c)] have been stored in the machine's memory.

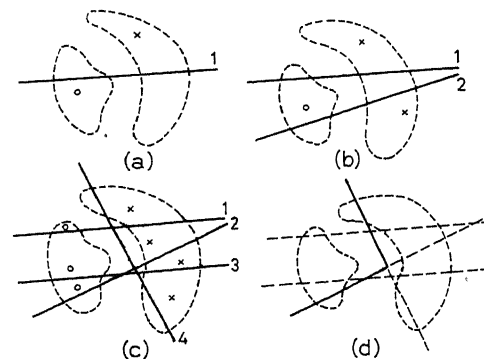


Figure 2

They break up the entire space into domains, in which any (and all) like points that have been shown are distributed, or else there is not a single indicated point at all. Following this, and in accordance with a predetermined algorithm, sections of the planes are 'erased' from the machine's memory, on both sides of which sections like points are distributed, or else those plane sections are erased in which points, shown during the teaching process, are distributed on one side, while on the other side, there are shown no points at all. As a result, a sectionally broken hypersurface is formed, which is made up of plane sections [Figure 2 (d)]. On different sides of this surface, points that belong to different patterns are distributed. The surface thus plotted is taken as the dividing one, and after that, upon being shown a new pattern element, i.e., upon the appearance of new points, the machine answers the question, 'What is this?', depending on which side of the surface the point lies.

In view of the fact that the planes are traced haphazardly each time, the repetition of the experiment will lead to the construction, not of the same, but of another surface of separation. This makes it possible to 'form parallel variants', and as a result of this, to increase the probability, to a large extent, of accurate replies, if during the teaching process there are constructed in parallel several surfaces of separation, and following that, the problem is allowed to be solved 'in accordance with the majority of votes'.

*Second algorithm—algorithm for construction of potential surfaces.* Upon being shown some point or other, the machine constructs a function of the point of receptor space, which attains a maximum in the point shown, and, remaining positive everywhere, decreases on being withdrawn from it in any direction. When several like points are shown, i.e., points that belong to one domain, the machine also constructs functions for each of these points, and afterwards combines them. It can easily be seen, that if the machine is shown several points that are distributed in a relatively uniform fashion through the domain, then,

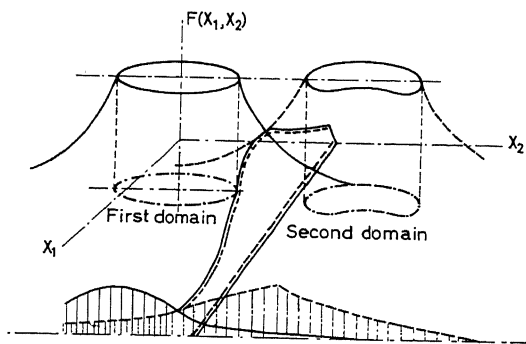


Figure 3

as a result, the machine constructs a surface, a distinctive potential function, which has high values ('hump') in domain points and drops sharply on being withdrawn from it. If two domains have to be separated, for example, then the machine constructs two such functions from the indicated points of these domains; one of them has a 'hump' over the first domain, and the second one over the second domain (Figure 3). In the case of the second algorithm, the teaching process is concluded with the construction of this potential function. Subsequently, upon being shown new points, the machine solves the question as to

which domain these points belong, in accordance with which one of the plotted potential functions has a greater value at this point. In the case of the second algorithm, the surface of separation is represented by the projection of the line of intersection of the plotted potential surfaces.

The author, together with Mr. Rozonoer and Mr. Braverman, proved some theorems connected with algorithm convergence. It was proved that using this algorithm in a finite number of steps it is possible to separate two domains if these are expressed in some simple conditions such as are always satisfied in practical problems. Furthermore the algorithm convergence will exist in spite of how one chooses the potential function as long as the function satisfies some simple conditions. Moreover, it is not necessary for the function to have the maximum in the shown point, to remain positive everywhere and to decrease on being withdrawn from shown point in any direction. These three conditions are more specific than necessary and are mentioned only to explain the algorithm more simply.

### Experiments on Teaching the Machine

In order to verify the compactness hypothesis and that of the two described algorithms, experiments were conducted on universal digital computers by E. M. Braverman (for the first algorithm) and O. A. Bashkirov (for the second algorithm).

In the first series of experiments (on the first algorithm), 160 outlines were prepared for each of the five figures—0, 1, 2, 3 and 5, i.e., 800 figure specimens altogether. For the purpose of instructional material, 40 specimens of each figure were selected, i.e., 200 figures, and 600 figures specimens were employed for subsequent machine testing. The results of these experiments are shown in the first column of Table 1.

It is evident, from the table, that in each variant the number of correct machine replies reached 89 per cent, and 'paralleling' seven variants, the number of correct replies comprised 98 per cent. Under these circumstances, it turned out, that at the end of the teaching process, there were taken up from 1,500 to 3,000 binary digit bits by the machine's memory.

In the second series of experiments (in reference to the algorithm for plotting potential functions), the operation was carried out both with five, as well as with all ten, figures simultaneously (see Table 1, fourth and fifth columns). In this case, four times less material was made use of for teaching purposes, i.e., for teaching purposes only 10 specimens of each figure were selected. On dividing five figures, in accordance with this algorithm, the machine yielded 100 per cent correct replies to 750 questions (using 150 specimens of each figure). During the experimental period, while teaching the machine simultaneous recognition of all 10 figures, 85 per cent correct replies were obtained to 1,400 inquiries. There is a clear understanding as to how to improve the algorithms, so that in this case, too, the number of correct replies might approach 100 per cent.

Experiments are being conducted, at the present time, on the utilization of the described algorithms for the purpose of teaching the machine the recognition of more complicated patterns; for example teaching the machine the simultaneous recognition of all 33 letters of the Russian alphabet in manuscript outline, the recognition of letters picked up by various typographic characters in the presence of 'obstacles' (smears, blurs, haphazard transposition of letters, etc.), the recognition of male and female portraits, images of the 'en face' and profile type, and so

forth. Despite the fact that these experiments have not yet been concluded, the successful results of the experiments with figures reassures us and gives us hope, that in more complex cases, too, it will be possible to obtain acceptable results.

Table 1

		First algorithm	Second algorithm	
			First experiment	Second experiment
Dividing Patterns		Figures 0, 1, 2, 3, 5	Figures 0, 1, 2, 3, 5	Figures 0, 1, 2, 3, 4, 5, 6, 7, 8, 9
Pattern elements Prepared	Of each figure Total number	160 800	160 800	150 1500
Selected for Instruction	Of each figure Total number	40 200	10 50	10 100
Used for Verification	Of each figure Total number	120 600	150 750	140 1400
Percentage of Correct Replies		in one experiment, 83-89 On 'parallel-ing' seven variants, 98.5	100	85

### Observations

*First Observation.* Along with the concept 'pattern', the concept 'association of patterns' can also be introduced. For example, the concepts 'small letter "a"' and 'capital letter "A"' are patterns, whereas the concepts 'any "a"' and '"A" letter' do not constitute a pattern. In fact, if a child, who does not know how to read, was shown only a row of outlines of the letter 'small "a"', and told what it was, and was then shown the letter 'capital "A"', and asked 'What is this?', then naturally, he would not give the correct reply. In order to teach the child (by the second teaching method) to recognize any letters, he must be shown 'small "a"s' as well as 'capital "A"s'. In such cases, it is convenient to speak of 'association of patterns'.

Both described algorithms are suitable for teaching the machine the recognition of pattern associations, if, in the teaching process, there are shown (approximately uniformly), the points of all the associated domains. Thus, for example, in utilizing the potential surfaces method, a 'two-hump' surface, rather than a 'one-hump' one will now be associated with the joining of the patterns 'any letter "a"' and '"A"'.

In fundamentally the same manner, it is possible to teach the machine to distinguish between any rectangles and triangles and any circles and ovals; between any figures and letters; between the portraits of three given faces and set of portraits of five other given faces, and so forth.

*Second Observation.* The described algorithms, in principle, always divide the patterns, but, of course, are not convenient in all cases, inasmuch as in some cases, for the purpose of division, too long a display may be required in the teaching process. The method is suitable in those cases where the domains, which

correspond to the patterns, are distributed in the receptor space at a good distance from one another. If, however, the patterns are such that the surface of separation must be 'strongly bent', then teaching by the described method will require the display of a very large number of points.

*Third Observation.* A good deal of attention is often devoted to having the machine distinguish correctly between the pattern elements that are projected on to various parts of the photo-field or pattern elements of different dimensions. For example, in the case of figure one in the left part of the field, in its upper part, the large and small pattern elements of figure one, and so forth, must be 'understood' by the machine as one and the same pattern element of the 'figure one'. From this paper's point of view, this entire problem has only a technical, and not a fundamental, significance. The centring and normalization of the pattern element is a problem that may readily be solved by other technical means, and which has no relation to the essence of the teaching process.

*Fourth Observation.* The compactness hypothesis made it possible not only to construct the described algorithms and to test them out on the machines, but also to understand the fundamental characteristics (peculiarities) of the experiments conducted by Rozenblatt, Steinbuch and others, with automatic machines specially constructed for this purpose. Thus, for example, it is possible to understand why the teaching of pattern recognition may be attained in the Perceptron or in Steinbuch's matrix only by adopting the compactness hypothesis, i.e., by an *a priori* assumption that the compact domains are divided in the receptor space.

*Fifth Observation.* It is natural to inquire: does the compactness hypothesis lie at the basis of at least the simplest processes in teaching pattern recognition to the living brain? This question steps outside the limits of the subject of this report and of the questions that are of interest to the International Federation of Automatic Control. For this reason, only brief mention is made of the physiological and psychological experiments being carried out, within the framework of this study, under the direction of Prof. S. N. Braines, at the Academy of Medical Sciences of the U.S.S.R.

Psychological experiments are being conducted with two series of blurrings. Each one of these series contains 150 blurs specially 'drawn' for these experiments by a machine, in such a way that all 150 blurs of the given series comprise a domain in the receptor space, and all 150 blurs of the other series comprise another domain.

During the experiments, the children are rapidly shown in turn the blurs from these two series, and they are faced with the problem of dividing them into two groups. In view of the fact that the blurs are deprived of any kind of meaningful significance for the children, that their outlines are haphazard and that they cannot be separated by any simple, immediately manifest signs, the successful conduct of this experiment would be a strong argument in favour of the belief that the compactness hypothesis in living organisms, too, lies at the basis of recognizing such a series of simple 'senseless' patterns.

Physiological tests, which are closely allied, in their concept, with the described psychological ones, are being carried out with rats and monkeys. Attempts are being made to develop reflexes, in animals, to the appearance of blurs from various series, and to judge from the behaviour of the animal as to the

degree to which it distinguishes between the blurs which are united or divided by only one property: in the receptor space, they are correlated with points that form various domains.

At the moment of writing this report, it is still too early to speak of the final results of these experiments, but the preliminary findings are reassuring.

### Teaching Recognition and Automatic Control

The success of the above-described experiments (see Table 1), as well as that of the experiments of Rozenblatt<sup>2</sup>, Selfridge<sup>3</sup>, Steinbuch<sup>4</sup>, Bongard<sup>5</sup> and others, makes it possible to foresee automatic control systems in the near future. So long as control is accomplished by a comparatively small number of regulating devices and the control purpose is to maintain the constancy of one or of a small number of interrelated magnitudes, the problems and possibilities of the learning systems are not considerable. But, in more complex control problems, where the control action is developed by taking into consideration a large number of readings (sometimes hundreds or thousands) of control-measuring devices, attempts to programme control processes in advance will inevitably be confronted with considerable difficulties. Control problems, under such conditions, become evident only if the possibility of teaching the system recognition of situations which arise at the input is kept in mind.

The difficulties confronted, under such conditions, are similar to those that originate in teaching even a qualified engineer the execution of the obligations involved in dispatching a complex undertaking. It makes no sense to specify, in advance, all the situations that might arise in the future, and, to work out instructions for the dispatcher's behaviour under all conditions. Experience indicates, however, that a dispatcher trainee, merely by observing the work of an experienced dispatcher for a period of time, and, under the influence of the simplest system of reward (encouragement), can make over this experience, i.e., he acquires the ability to come up with correct solutions even in

new situations which had not been observed by him earlier in the teaching process.

How does this take place? To what degree are the processes which occur in this situation similar to those that were attempted on the models in our experiments? At the present time, it is difficult to answer these questions, and it is not clear as to what this study, and studies similar to it, yield towards obtaining substantiated replies. However, the value of such experiments to the technology of automatic control is evident, inasmuch as they make it possible to hope for a substantial expansion in the possibilities involved in control systems. Up to the present time, in order to carry out various control operations, the systems had to be constructed differently beforehand, or else, had to have different programmes. At the present time, it is clear that a universal control system with a multi-purpose teaching programme will be capable of being taught the execution of the most varied control problems. The hope has emerged that control systems in the future, rather than being constructed or programmed in different ways, will be capable of being taught in different ways and 'educated' in different ways. It is difficult to evaluate the possibilities that this may open for automatic control technology, if this term is to be understood within any broad framework of design.

### References

- 1 BRAVERMAN, E. M. Experiments on teaching machines the recognition of visible patterns. *Automat. Telemekh.* XXIII, No. 3 (1962)
- 2 ROZENBLATT, F. Perceptron simulation experiments. *Proc. IRE.* 48, No. 3 (1960)
- 3 SELFRIDGE, O. G. Pandemonium: a paradigm for learning. *Symp. Mechanization of Thought Process.* England (November, 1958)
- 4 STEINBUCH, K. Automatische Zeichenerkenntnis. *Nachrichtentechnische Zeitschrift*, II (1958), H. 9/5
- 5 BONGARD, M. M. Modelling the recognition process on a digital computer. *Biophys.* VI, No. 2 (1961)
- 6 HAY, J. C. and others. The Mark I Perceptron—design and performance. *IRE Internat. Conv. Rec.* (1960), 2

### DISCUSSION

P. NASLIN, *Laboratoire Central de l'Armement, Arcueil (Seine), France*

On page 4 of the paper, the problems of centring and normalizing are disregarded, being easily solvable by other means and having no relation to the essence of the learning process. Does this statement hold for the following problems, which are essentially of the same nature?

(1) How can the machine properly orient a character before it has identified it?

(2) If the machine is faced with a small character attached to a large blur, how can it eliminate the meaningless blur before it has identified the meaningful character?

While those problems may be of secondary importance for the technical applications of pattern-recognition machines, they seem to be relevant to the learning process considered *per se*, especially as they present little difficulty to the human eye and brain.

M. A. AIZERMAN, *in reply*

The objective of our work was not the development of reading machines but rather the study of fundamental problems in machine imitation of the learning process. We used the recognition of visual

patterns only as an example of the learning process. Naturally we tried to disregard problems connected only with the recognition of visual patterns and not with the learning of other kinds of pattern recognition, e.g. of recognition of sounds or with development of machines for medical and technical diagnostic purposes. As for one of the internal problems of reading automata, the centring problem, this can be understood either as an independent technical problem or can be solved in the course of and in connection with the recognition process. However, these problems were outside the limits of my report.

T. KASVAND, *Department of Electronics, University of Southampton, Southampton, England*

It is evident that a superior system can 'understand' the operation of an inferior system. However, it has been said that this superior system cannot understand its own method of operation, since that would require more information than the system is able to store and handle. If there is any basis to this belief there may exist an insurmountable gap between the human intellect and the intelligent systems that we are trying to devise. Do you feel that this fear is ungrounded, or that the speed of computers will compensate for the lack of deeper understanding?

Is there but one mathematical operation that maps the responses of the  $n$  receptors to a point in the  $n$ -dimensional space, or is there a whole sequence of operations, each in turn dealing with increasingly more complex concepts, where the final operation produces the single point in the  $n$ -dimensional space?

Has any attempt been made to transform the  $n$ -dimensional space to a suitably defined 3-dimensional space? This may throw some light on the way the brain handles the recognition problem.

Since there is no well-defined terminology I regret that some of my questions may not be clear.

M. A. AIZERMAN, *in reply*

In his first question Dr. Kasvand touched upon the principal problems of possibilities of machines which I intentionally avoided in my report. Up to now, to programme for any particular process, a human operator had to know how to realize this process and also to understand how it could be realized, that is, he had to be able to compose the algorithm—the programme of this process. The idea of the report was to go beyond this and to show that the machine can be taught to do the same operation that a man can perform even if he himself does not understand how he does it (by demonstration). It seems to me that this is a way to expand the machine possibilities, and from my point of view it is more important to advance in this direction than to try to foresee the limits which can restrict progress along this line. Any attempt to find such limits requires a more precise statement of the problem than we can make at present.

On the second question I would like to explain that as  $n$ -dimensional receptor space we use directly the space of receptor states without any intermediate signal transformations, and, in particular, without decrease of space dimension in comparison with the number of receptors. The concept 'compactness of pattern' is pointless if the space is not fixed. The meaning of the compactness hypothesis consists in the fixing of this space: namely, we proceed from the assumption that the pattern is a collection of points composing a region in the space of receptors. Of course, we can choose transformations which make this region 'more compact', e.g. just convex, and in this way we can make easier the problem of division of this region. However, there are transformations which generally disturb the compactness. The problem of distinguishing 'good transformers' requires further research or new ideas.

J. E. GIBSON, *Control and Information Systems Laboratory, Purdue University, Lafayette, Indiana, U.S.A.*

I listened with interest to this preliminary report of the work on learning systems which Professor Aizerman's laboratory is beginning. The 'compactness hypothesis', which is advanced without proof, apparently implies that every pattern constitutes a compact set, that is, it can be isolated by hypersurfaces from every other pattern of the domain considered. This notion has been previously assumed by Rozenblatt, Widrow, Matson and others, but it appears not to have been stated explicitly previously.

The first algorithm of Aizerman for separating patterns by piecewise linear hyperplanes has been employed by Widrow's Madeline among others. The second algorithm, that of the potential hill, appears to be new and is an interesting contribution. It appears to have the drawback, however, that a digital computer will be required to realize it. This adjustment is definitely non-trivial since the shape of the potential function will affect the accuracy of recognition and its convergence. All of these matters remain to be discussed in the future.

Finally, I feel that the application of learning theory to the solution of automatic control problems will be most fruitful in an on-line extension of current adaptive control theory rather than in the off-line training sequence approach suggested by Professor Aizerman.

We look forward with great interest to future detailed reports of the progress of Professor Aizerman and his group in this very difficult research area.

M. A. AIZERMAN, *in reply*

I am very grateful to Professor Gibson for his interest in our work, but I cannot agree that our algorithm of random planes concurs with that realized by Widrow in his Madeline. The Madeline algorithm slightly resembles ours only in appearance.

The principal feature of our algorithm is that the random planes are built as the points are appearing and that the piecewise surface is formed by erasing superfluous parts of these frames. As I understand it, Widrow's algorithm does not contain such a process.

# On the Theory of Self-tuning Systems with a Search of Gradient by the Method of Auxiliary Operator

I. E. KAZAKOV and L. G. EVLANOV

## Summary

The paper considers the method of auxiliary operator for the determination of the gradient of index of the quality of control with respect to tuned parameters for continuous self-tuning. A theoretical investigation is given of the dynamics and precision of the extremal tuning of parameters of an automatic control system with the utilization of a method of auxiliary operator. The effectiveness of the considered system of self-tuning with random disturbances is demonstrated.

## Sommaire

Ce rapport présente la méthode de l'opérateur auxiliaire pour déterminer le gradient de l'indice de qualité de commande en ce qui concerne les paramètres accordés pour un accord automatique continu. On examine en particulier l'aspect théorique de la dynamique et la précision de l'accord extrême des paramètres avec cette méthode dans le cas d'une commande automatique. On démontre l'efficacité de cette méthode dans le cas de perturbations aléatoires.

## Zusammenfassung

Dieser Beitrag behandelt den Einsatz eines zusätzlichen Operators zur Gradientenbestimmung des Index der Regelgüte für die angepaßten Parameter bei kontinuierlicher Selbstanpassung. Die Dynamik und die Genauigkeit der Extremwertanpassung solcher Systeme mit Hilfsoperator werden theoretisch untersucht. Die Wirksamkeit eines derartigen Systems mit Selbstanpassung bei Zufallsstörungen wird dargestellt.

## Structure and Equations of a Self-tuning System

In many cases important in practice, automatic control systems may be represented in the form of a generalized system illustrated in Figure 1. The controlled plant is characterized by an operator of a given structure  $A(\eta)$ , where  $\eta$  is a group of parameters for which *a priori* information is lacking. The system of control is described by an operator  $B(\xi)$  which depends on the group of parameters  $\xi_i$  ( $i = 1, 2, \dots, n$ ) which may be adjusted.

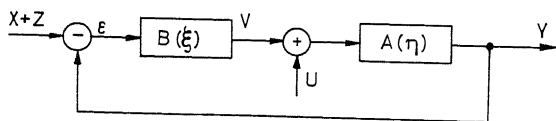


Figure 1

In actual systems, the entirety of values of each parameter  $\xi_i$  forms a finite multitude  $\Xi_i$ . The input signals of the system are  $X(t)$ , the useful random signal, and  $Z(t)$ ,  $U(t)$ , random disturbances.

The equations of the automatic control system are as follows:

$$\begin{aligned} Y &= A(\eta) [V + U] \\ V &= B(\xi) \varepsilon \\ \varepsilon &= X + Z - Y \end{aligned} \quad (1)$$

In order to assure high quality operation of the automatic control system it is necessary to achieve tuning of parameters of the operator  $B(\xi)$  in the presence of variation of the characteristics of the input useful signal  $X(t)$ , of the characteristics of disturbances  $Z(t)$ ,  $U(t)$ , and also in the presence of variation of parameters  $\eta$  of the operator of the controlled plant.

In order to construct a circuit for self-tuning, an index of quality  $I$  of the automatic control system is introduced. The index of quality  $I$  is a function, or in the general case it is a functional of tuned parameters. Ordinarily the index of quality  $I$  is computed on the basis of error  $\varepsilon$  of the system:

$$I = Nf(\varepsilon, \xi) \quad (2)$$

where  $N$  is an operator or a functional,  $f(\varepsilon, \xi)$  is a function of the error of the system depending upon the error  $\varepsilon$  and the tuned parameters  $\xi$ .

In order to adjust the parameters of the system use is made of the broad possibilities offered by the method of steepest descent or gradient, a discussion of which is considered by Feldbaum<sup>1</sup>. Applying this method for tuning parameters  $\xi$  one has:

$$\dot{\bar{\xi}} = \lambda \text{grad } I \quad (3)$$

where  $\lambda$  is the scalar multiplier, and  $\bar{\xi}$  is a vector function of the velocities of tuned parameters. In accordance with the gradient method the self-tuning system assures the tuning of parameters  $\xi$  for the optimal value of index of quality  $I_0$ . In the general case

$$I_0 = \inf_{\xi_i \in \Xi_i} I(\xi) \quad \text{or} \quad I_0 = \sup_{\xi_i \in \Xi_i} I(\xi) \quad (4)$$

In the particular case when the lower (upper) boundary of the multitude  $\xi_i$  is attained within  $\Xi_i$ ,

$$I_0 = \text{extremum } I(\xi) \quad (5)$$

For a complete description of the circuit for self-tuning it is necessary to determine the method of computation of the components of the gradient from the quality index for the tuned parameters. In the given investigation a method is applied which, in the following is termed the method of an auxiliary operator. Its essence consists of the following.

If the information on operators  $B(\xi)$  and  $A(\eta)$  is known *a priori*, it is possible to construct a certain auxiliary operator  $C(\xi, \eta)$  whose application to the error of the tracking system makes it possible to compute the components of the gradient vector.

The derivative  $\partial I / \partial \xi_i$  is computed by the direct differentiation of the expression (2) assuming that the operators  $N$  and differentiations with respect to  $\xi_i$  are commutative.

$$\frac{\partial I}{\partial \xi_i} = N \frac{\partial f(\varepsilon, \xi)}{\partial \varepsilon} \cdot \frac{\partial \varepsilon}{\partial \xi_i} + N \frac{\partial f(\varepsilon, \xi)}{\partial \xi_i} \quad (6)$$

The derivative  $\partial \varepsilon / \partial \xi_i$  will be calculated by differentiating the system of eqns (1). The derivative of the error  $\varepsilon$  with respect to  $\xi_i$  is equal to

$$\frac{\partial \varepsilon}{\partial \xi_i} = - \frac{\partial Y}{\partial \xi_i} \quad (7)$$

since the input signals  $X(t)$ ,  $Z(t)$  do not depend upon  $\xi_i$ . The derivatives of the output signal are computed:

$$\frac{\partial Y}{\partial \xi_i} = A(\eta) \frac{\partial B(\xi)}{\partial \xi_i} \varepsilon + A(\eta) B(\xi) \frac{\partial \varepsilon}{\partial \xi_i} \quad (8)$$

Excluding from (7) and (8)  $\partial Y / \partial \xi_i$  and transforming, one obtains:

$$\frac{\partial \varepsilon}{\partial \xi_i} = - [1 + A(\eta) B(\xi)]^{-1} A(\eta) \frac{\partial B(\xi)}{\partial \xi_i} \quad (9)$$

Introducing the designation

$$C_i = [1 + A(\eta) B(\xi)]^{-1} A(\eta) \frac{\partial B(\xi)}{\partial \xi_i} \quad (10)$$

one writes:

$$\frac{\partial \varepsilon}{\partial \xi_i} = - C_i(\eta, \xi) \varepsilon \quad (11)$$

or

$$\text{grad } \varepsilon = - \bar{C}(\eta, \xi) \varepsilon \quad (12)$$

where  $\bar{C}(\eta, \xi)$  is an auxiliary operator-vector which is completely determined by the operators  $A(\eta)$ ,  $B(\xi)$ . Thus, the gradient of the quality index for tuned parameters is determined by eqns (6) and (11).

The method of auxiliary operator requires an *a priori* knowledge of information on the system, and this somewhat restricts its generality. However, there exists in technology an area of applicability of the method inasmuch as the predominant majority of created automatic control systems can be described mathematically.

The advantages of the method are the absence of trial load changes and the possibility of accelerating and simplifying the process of computation of the gradient components. In self-tuning systems with a search of gradient by the method of trial load changes, *a priori* information on the plant, other than the knowledge of the band pass of the system, is not required. This permits a correct selection of the frequency of the trial load changes and constitutes the advantage of this method. However, its basic shortcoming is the limited quick response imposed by the finite band pass width of the system. In the considered method the band pass of the mathematical model of the system (operator  $C$ ) may be artificially broadened by changing the time scale of the solution. The possibility of simplifying the process of computation is based on the substitution for a complex operator  $C$  of an approximate and simpler expression.

The auxiliary operator  $\bar{C}(\eta, \xi)$  depends upon the parameters of the plant and the system of control. A typical case

is one of absence of *a priori* information on parameters  $\eta$ . Information on parameters of the plant may be obtained on the basis of application of a tracking system, certain aspects of whose application were considered by Margolis and Leondes<sup>2,3</sup>.

The structure of the operator of model  $A(\zeta)$  is based on the utilization of *a priori* information on the plant. The entirety of parameters  $\zeta$  of the operator of the model is tuned for the value  $\eta$ . The circuit of the tracking model is constructed quite analogously to the circuit for tuning. Introducing an index of approximation  $J$  of parameters  $\zeta$  into parameters  $\eta$ ,

$$J = L\phi(\varepsilon_1) \quad (13)$$

where  $L$  is an operator for computing the index  $J$ , and  $\phi(\varepsilon_1)$  is a function of the error. The error is determined by the relationship

$$\varepsilon_1 = Y_M(t) - Y(t) \quad (14)$$

Here  $Y_M(t)$  is an output signal of the model determined by the expression

$$Y_M(t) = A(\zeta) V \quad (15)$$

The change of the parameters of the model is carried out by the method of steepest descent:

$$\bar{\zeta} = \lambda_1 \text{grad } J \quad (16)$$

where  $\lambda_1$  is a scalar multiplier, and  $\bar{\zeta}$  is a vector function of the velocities of the tuned parameters of the model.

In order to determine the components of the gradient one applies the method of auxiliary operator:

$$\frac{\partial J}{\partial \zeta_i} = L \frac{\partial \phi(\varepsilon_1)}{\partial \varepsilon_1} \cdot \frac{\partial \varepsilon_1}{\partial \zeta_i} \quad (17)$$

Differentiating the relationship (14) with respect to  $\zeta_i$ , one has:

$$\frac{\partial \varepsilon_1}{\partial \zeta_i} = \frac{\partial Y_M}{\partial \zeta_i} = \frac{\partial}{\partial \zeta_i} A(\zeta) V = \frac{\partial A(\zeta)}{\partial \zeta_i} V \quad (18)$$

hence it follows that the auxiliary operator in a given case is an operator-vector  $\bar{G}(\zeta)$  with components

$$G_i(\zeta) = \frac{\partial A(\zeta)}{\partial \zeta_i} \quad (19)$$

Thus

$$\text{grad } J = L \left\{ \frac{\partial \phi(\varepsilon_1)}{\partial \varepsilon_1} \bar{G}(\zeta) V \right\} \quad (20)$$

Equations (13), (14), (15), (16) and (20) describe the operation of the tracking model. A useful output of the circuit of the model is the entirety of parameters of model  $\zeta$ . For ideal operation of the model  $\zeta \equiv \eta$ . An actual model assures the attainment of parameters  $\zeta$  close to values  $\eta$ , and therefore, strictly speaking, in the operator  $\bar{C}$  it is necessary to replace parameters  $\eta$  by  $\zeta$ .

The complete block diagram of the self-tuning system in accordance with eqns (1), (3), (6), (11), (14), (15), (16) and (20) is presented in Figure 2. The schematic diagram was proposed by Evlanov.

The structure of the self-tuning system contains three circuits: the basic circuit of the system, the circuit of the tracking



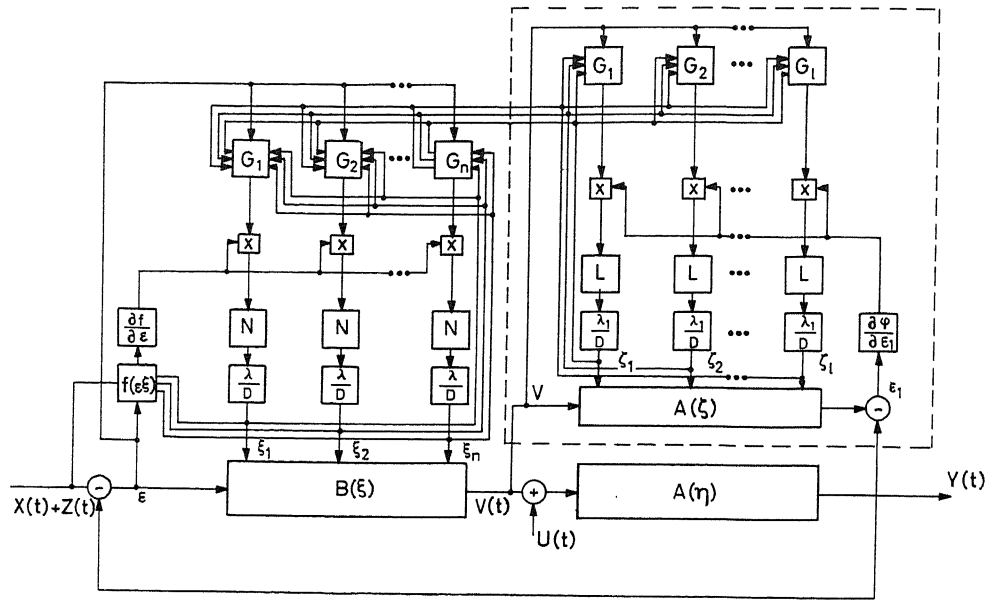


Figure 2

model, and the circuit for tuning the parameters. The circuit of the tracking model assures the reception of information on the parameters of the operator of the plant. In the following the operation of the circuit of the tracking model is assumed to be ideal, that is,  $\zeta \equiv \eta$ . The circuit for tuning the parameters assures the tuning of parameters of the control system in accordance with the given optimal value of the quality index of the system.

#### Investigation of a Self-tuning System in a Quasi-stationary Regime

A typical regime of operation of a self-tuning system is the case of a change of parameters  $\eta$  of the operator  $A(\eta)$  of the plant and of the characteristics of external random disturbances  $X, Z, U$  which are slow compared with the duration of transient processes in the basic circuit of the system. In this case it is permissible to consider the circuits for tuning parameters and the tracking model on the one hand, and the basic circuit on the other hand, as being autonomous, since the tuned parameters  $\xi$  and parameters  $\eta$  may be considered as constant during the time of process control in the basic circuit. It is also assumed that the tracking model carries out its functions in an ideal manner. Under these conditions the process of self-tuning of parameters  $\xi$  of operator  $B(\xi)$  is investigated in the vicinity of extremum of the quality index  $I$ .

The presence of extremum in the quality index  $I$  of the system with respect to all or several of the tuned parameters is an important property of self-tuning systems which permits them to be tuned for an optimal regime. If the error of the system  $\varepsilon$  or another characteristics does not possess extremal properties, then it is possible to construct an extremal quality index by artificial means depending upon the direction of the target of the automaton. This will be shown below by an example of a typical tracking system. For the time being, however, it is assumed that the quality index  $I$  possesses extremal properties.

The random error  $\varepsilon$  of the basic circuit can be expressed in the form

$$\varepsilon = m_\varepsilon + \varepsilon^0 \quad (21)$$

where  $m_\varepsilon$  is the mathematical expectation, and  $\varepsilon^0$  is the centring component of magnitude  $\varepsilon$ . In the function of the error  $f(\varepsilon, \xi)$  we shall also factor out the mathematical expectation

$$f(\varepsilon, \xi) = Mf(\varepsilon, \xi) + f^0(\varepsilon, \xi) \quad (22)$$

where  $M$  is the operation of mathematical expectation,  $f^0(\varepsilon, \xi)$  is the random centred component.

The quality index of control  $I$  introduced previously may now be presented as:

$$I^* = NI^* + Nf^0(\varepsilon, \xi) \quad (23)$$

where the designation  $I^*$  is introduced for the statistical quality index of control

$$I^* = Mf(\varepsilon, \xi) \quad (24)$$

Computing the components of the gradient of the quality index of control by parameters  $\xi_i$ , one obtains:

$$\frac{\partial I}{\partial \xi_i} = N \frac{\partial I^*}{\partial \xi_i} + N \frac{\partial f^0}{\partial m_\varepsilon} \cdot \frac{\partial m_\varepsilon}{\partial \xi_i} + N \frac{\partial f^0}{\partial \varepsilon^0} \cdot \frac{\partial \varepsilon^0}{\partial \xi_i} + N \frac{\partial f^0}{\partial \xi_i} \quad (25)$$

Representing the statistical quality index  $I^*$  of control in the vicinity of the investigated extremum by a quadratic form in terms of deviations  $u_i = \xi_i - \xi_{i0}$  of parameters  $\xi_i$  from the optimal values  $\xi_{i0}$ , and considering that

$$\left[ \frac{\partial I^*}{\partial \xi_i} \right]_{\xi_i = \xi_{i0}} = 0$$

at the point of extremum, we shall obtain for the current values of  $\partial I^* / \partial \xi_i$  the expressions:

$$\frac{\partial I^*}{\partial \xi_i} = \sum_{j=1}^n \frac{1}{2} \left[ \frac{\partial^2 I^*}{\partial \xi_i \partial \xi_j} \right]_0 u_j \quad (26)$$

Differentiating expressions (24) twice with respect to parameters  $\xi_i, \xi_j$  and utilizing a system of equations of the basic



circuit of control for optimal parameters  $\xi_{i0}$  of operator  $B(\xi)$ , one computes the coefficients

$$\left[ \frac{\partial^2 I^*}{\partial \xi_i \partial \xi_j} \right]_0$$

in the form:

$$\begin{aligned} \left[ \frac{\partial I^*}{\partial \xi_i \partial \xi_j} \right] = M \left\{ \frac{\partial^2 f(\varepsilon_0, \xi_0)}{\partial \varepsilon_0^2} (C_{j0} \varepsilon_0) (C_{i0} \varepsilon_0) \right. \\ \left. + \frac{\partial f(\varepsilon_0, \xi_0)}{\partial \varepsilon_0} (C_{j0} C_{i0} \varepsilon_0) + \frac{\partial^2 f(\varepsilon_0, \xi_0)}{\partial \varepsilon \partial \xi_j} (C_{i0} \varepsilon_0) \right. \\ \left. + \frac{\partial^2 f(\varepsilon_0, \xi_0)}{\partial \xi \partial \xi_i} (C_{j0} \varepsilon_0) + \frac{\partial^2 f(\varepsilon_0, \xi_0)}{\partial \xi_i \partial \xi_j} \right\} \quad (27) \end{aligned}$$

where  $C_{i0}(\xi_0, \eta)$  are the auxiliary operators (10) for optimal values of parameters  $\xi_{i0}$ .

Introduce the designations:

$$\frac{1}{2} \left[ \frac{\partial^2 I^*}{\partial \xi_i \partial \xi_j} \right]_0 = a_{ij} \quad (28)$$

Taking into account also that

$$\frac{\partial m_\varepsilon}{\partial \xi_i} = -C_1 m_\varepsilon, \quad \frac{\partial \varepsilon^0}{\partial \xi_i} = -C_i \varepsilon^0 \quad (29)$$

the formula (25) is written for the components of the gradient of the magnitude  $I$  in the form:

$$\frac{\partial I}{\partial \xi_i} = N \sum_{j=1}^n a_{ij} u_j - N \frac{\partial f^0}{\partial m_\varepsilon} C_i m_\varepsilon - N \frac{\partial f^0}{\partial \varepsilon_i^0} C_i \varepsilon^0 + N \frac{\partial f^0}{\partial \xi_i} \quad (30)$$

Substituting the expression (30) into formula (3), one obtains a system of equations of the circuits for tuning the parameters  $\xi_i$  in a scalar form:

$$\dot{\xi}_i = \lambda N \sum_{j=1}^n a_{ij} u_j - \lambda N \frac{\partial f^0}{\partial m_\varepsilon} C_i m_\varepsilon - \lambda N \frac{\partial f^0}{\partial \varepsilon_i^0} C_i \varepsilon^0 + \lambda N \frac{\partial f^0}{\partial \xi_i} \quad (31)$$

From this one obtains a system of linear equations for the determination of mathematical expectations of deviations  $m_{u_i}$  of tuned parameters from the optimal values:

$$\dot{m}_{u_i} - \lambda N \sum_{j=1}^n a_{ij} u_j = -\dot{\xi}_{i0} \quad (32)$$

In order to determine random components of deviations of tuned parameters  $u_i^0$  one obtains the following system of linear equations:

$$\begin{aligned} \dot{u}_i^0 - \lambda N \sum_{j=1}^n u_j^0 a_{ij} = -\lambda N \left[ \frac{\partial f^0}{\partial m_\varepsilon} \right] C_{i0} m_{\varepsilon_0} \\ - \lambda N \left[ \frac{\partial f^0}{\partial \varepsilon_0^0} \right] C_{i0} \varepsilon_0^0 + \lambda N \frac{\partial f^0}{\partial \xi_i} \quad (33) \end{aligned}$$

An analysis of approximate linear equations (32) makes it possible to evaluate the stability of the process and to determine the systematic errors of self-tuning of parameters  $\xi_i$ . In particular, if the basic circuit of control is stationary and possesses astatism of the  $k$ th order, then for stationary random distur-

ances  $Z$  and  $U$ , and for an additive component of the useful signal  $X$  in the form of a polynomial of the  $k$ th order, the left-hand parts of eqns (32) are stationary. In this widely encountered case the stability of self-tuning of the parameters is characterized by the properties of the characteristic equation. In this case the investigation of stability is carried out by ordinary means. In the general case the systematic components of the errors of parameters are computed by equations:

$$m_{u_i}(t) = - \sum_{j=1}^n \int_0^t g_{ij}(t, \tau) \dot{\xi}_{j0}(\tau) d\tau \quad (34)$$

where  $g_{ij}(t, \tau)$  are the weighting functions of the system of eqns (32). If  $\xi_{i0} = \text{const.}$ , then the systematic values of errors of tuning of parameters  $m_{u_i} = 0$ . Dispersions of the errors of parameters are determined on the basis of the system of eqns (33) by applying the theory of transformation of random functions<sup>4</sup>.

From the analysis of stability, duration of transient processes of tuning, and evaluation of the precision, one chooses the coefficient  $\lambda$  and also other characteristics of the tuning circuits.

The final evaluation of mathematical expectation of the error in the basic circuit of the system under the action of self-tuning circuits is obtained by the formula:

$$m_\varepsilon = m_{\varepsilon_0} + \sum_{i=1}^n m_{\varepsilon_i} \quad (35)$$

where  $m_{\varepsilon_0}$  is the mathematical expectation of the control error of control  $\varepsilon$  for an optimal value of parameters.

The magnitudes  $m_{\varepsilon_i}$  are determined by the expressions:

$$m_{\varepsilon_i} = C_{i0}(\xi_0) [b_i m_{u_i}]$$

where  $C_{i0}(\xi_0)$  are the auxiliary operators for optimal values of parameters  $\xi_{i0}$  and the magnitudes  $b_i$  are equal to

$$b_i = \left[ \frac{\partial B(\xi)}{\partial \xi_i} \right]_0 m_{\varepsilon_0} \quad (36)$$

The evaluation of dispersion of the error in the basic circuit is computed by the formula:

$$D_\varepsilon = D_{\varepsilon_0} + 2 \sum_{i=1}^n k_{\varepsilon_0 \varepsilon_i} + \sum_{i,j=1}^n k_{\varepsilon_i \varepsilon_j} \quad (37)$$

where  $D_{\varepsilon_0}$  is the dispersion for optimal values of parameters  $\xi_{i0}$ ,  $K_{\varepsilon_0 \varepsilon_i}$ ,  $K_{\varepsilon_i \varepsilon_j}$  are the coefficients of correlation of random components of the control error  $\varepsilon_i^0$ , and the magnitudes  $\varepsilon_i^0$  are equal to

$$\varepsilon_i^0 = -u_i^0 [C_{i0}(\xi_0) m_{\varepsilon_0}] \quad (38)$$

### Linear Tracking System with One Tuned Parameter

The application of the method to a linear tracking system, with one tuned parameter, is now described. In tracking systems, as a rule, the index of control quality is assumed to be the second initial moment of error  $\varepsilon$ . This magnitude does not possess extremal properties with respect to parameters  $\xi$  corresponding to the change of input random actions  $X$ ,  $Z$ ,  $U$ .

Now consider an example of a tracking system having the following characteristics:  $A(\eta) = \frac{\eta}{D}$ ,  $B(\xi) = \xi_1$ ,  $X = at$ ,  $U = 0$ ,

$$m_z = 0, s_z = \frac{D_z}{\pi} \frac{\beta}{(\omega^2 + \beta^2)}$$

and values of parameters given by  $\eta_1 = 10$ ,  $a = 0, 1$ ,  $D_z = 10^{-4}$ ,  $\beta = 100$ . The second initial moment of error  $\varepsilon$  in a stabilized regime is equal to:

$$\alpha_\varepsilon = \frac{a^2}{\xi_1^2 \eta_1^2} + \frac{D_z \beta}{\xi_1 \eta_1 + \beta} \quad (39)$$

This relationship has no extremum with respect to parameter  $\xi_1$ .

In the theory of optimal filtration the magnitude  $\varepsilon^* = \varepsilon - Z = X - Y$  is considered as an error. The second initial moment of this magnitude possesses extremal properties. Thus, under the conditions of the preceding example the magnitude  $\alpha_\varepsilon^*$  is equal to:

$$\alpha_\varepsilon^* = \frac{a^2}{\xi_1^2 \eta_1^2} + \frac{D_z \eta_1 \xi_1}{\xi_1 \eta_1 + \beta} \quad (40)$$

This function has an extremum with respect to parameter  $\xi_1$ .

It is possible to measure directly the magnitude  $\varepsilon^*$  in tracking systems using *a priori* information on the statistical properties of the input useful signal and the disturbances. In practice it is possible to measure the error  $\varepsilon$  and the signal  $Z_1 = Z_1(X, Z)$  related to  $Z$ . For instance, the function  $Z_1$  may be obtained by filtering with special filters the input signal  $X + Z$  and utilizing the information as the spectrum of the frequencies of the disturbance  $Z$ , as a rule, is substantially broader than the spectrum of the useful signal  $X$ . Then the function  $Z_1$  will possess characteristics which are close to the characteristics of the function  $Z$ .

Measuring the magnitudes  $\varepsilon$  and  $Z_1$  it is possible to formulate artificially a quality index having an extremal characteristic with respect to the gain  $\xi_1$  of the correcting circuit  $B(\xi)$ . For this the function of the error is assumed to have the form:

$$f(\varepsilon, \xi) = \varepsilon^2 + \psi(\xi_1) Z_1^2 \quad (41)$$

The function  $\psi(\xi_1)$  may be chosen in a specific case, for instance, from the condition of proximity of the extrema of functions  $M[\varepsilon - Z]^2$  and  $M[\varepsilon^2 + \psi(\xi_1) Z^2]$  with respect to parameter  $\xi_1$  for a statistically prescribed input signal.

As an illustration of the method of prescribing a function  $\psi(\xi_1)$  let us consider the case of good filtering when it is possible to neglect the component  $X$  in function  $Z_1$ . Let us determine  $\psi(\xi_1) = \nu \xi_1$ , where  $\nu$  is a constant coefficient computed from the condition of proximity of the values of parameters  $\xi_{10}$  for extremal values of the functions  $\tilde{\alpha} = M\varepsilon^2 + \nu \xi_1 D_z$  and  $\alpha_\varepsilon^* = M(\varepsilon - Z)^2$ .

In Figure 3 there are presented graphs of functions  $\tilde{\alpha}$  and  $\alpha_\varepsilon^*$  corresponding to the minimal value and computed for the preceding example. For  $\nu = 0.1$  the minima of the functions (curves with an index 1) coincide closely, and the optimal value of parameter  $\xi_{10} = 3.0$ . The change in a sufficiently broad range of probability characteristics of disturbance  $Z$ , useful signal  $X$ , and parameter  $\eta$  leads to a distortion of the form of the curves  $\tilde{\alpha}$  and  $\alpha_\varepsilon^*$ . However, their minima coincide, but are not reached for other values of parameter  $\xi_{10}$  as shown in Figure 3. In Figure 3 the index 2 denotes curves for  $D_z = 10^{-3}$  and the previous values of other parameters.

In Figure 4 there is shown a schematic diagram of a linear tracking system with tuning of the gain  $\xi_1$  for  $\psi(\xi_1) = \nu \xi_1$ . The function  $Z_1$  is separated with the aid of a band pass filter or a filter of high frequencies. Then the signal is supplied to a square wave generator and a circuit with gain  $\nu \xi_1$ , and then

to a low frequency filter. Now consider the quasi-stationary regime of self-tuning of parameters. Eqn (31) of tuning of parameter  $\xi_1$  stated with respect to deviation  $u_1$  assumes the form:

$$[(TD + 1)D - \lambda a_1] u_1 = -2 \lambda m_{\varepsilon_0} [C_{10}(0) + C_{10}(D)] \varepsilon_0^0 - D \xi_0 + 2 \lambda \nu m_z Z_1^0 \quad (42)$$

where

$$a_1 = M \{ [C_{10}(D) \varepsilon_0^0]^2 + [\varepsilon_0 (C_{10}^2(D) \varepsilon_0^0)] \} > 0 \quad (43)$$

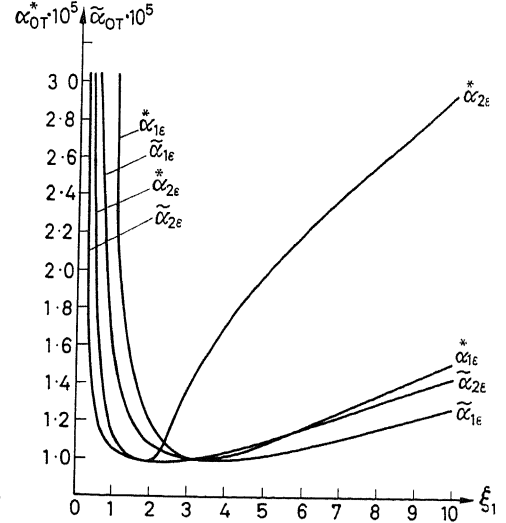


Figure 3

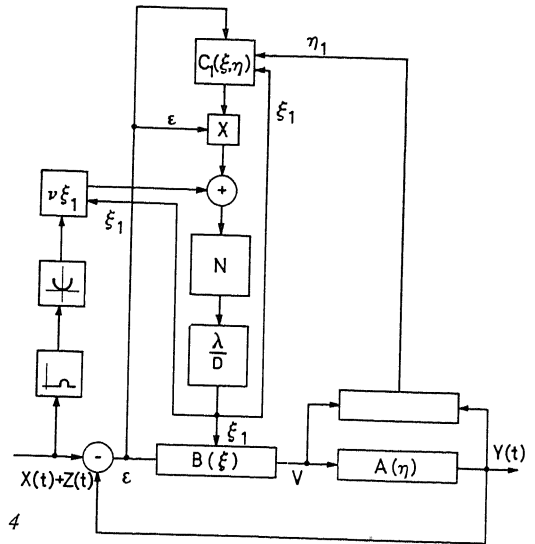


Figure 4

From these one obtains the following equation for the determination of mathematical expectation  $m_{u1}$ :

$$(TD^2 + D - \lambda a_1) m_{u1} = -D \xi_{10} \quad (44)$$

For  $\lambda < 0$  the stable process of tuning is assured. When one determines the centred random component  $u_1^0$ , one obtains the equations:

$$[TD^2 + D - \lambda a_1] u_1^0 = -2 \lambda m_{\varepsilon_0} [C_{10}(0) + C_{10}(D)] \varepsilon_0^0 + 2 \lambda \nu m_z Z_1^0 \quad (45)$$

The magnitude  $m_{z_1}$  may be set equal to zero by proper selection of the corresponding filter. Taking this into account and also utilizing expressions for  $\varepsilon_0^0$  in terms of  $X^0 + Z^0$ , one obtains from eqn (45)

$$u_1^0 = \Phi_1(D)(X^0 + Z^0) \quad (46)$$

where

$$\Phi_1(D) = \frac{-2\lambda m_{e_0}[C_{10}(0) + C_{10}(D)]}{(TD^2 + D - \lambda a_1)[1 + A(D)B_0(D)]} \quad (47)$$

In this case, for computing the dispersion of parameter  $u_1$  in a stabilized regime, one obtains:

$$D_{u_1} = \int_{-\infty}^{\infty} |\Phi_1(i\omega)|^2 [S_x(\omega) + S_z(\omega)] d\omega \quad (48)$$

where  $S_x$  and  $S_z$  are the spectral densities of random functions  $X$  and  $Z$ . For  $\xi_{10} = \text{const.}$  the magnitude  $m_{e_0} = 0$  in the stabilized regime. In this case the systematic error of a tracking system with self-tuning in a stabilized regime of operation is equal to  $m_e = m_{e_0}$ , that is, equal to the systematic error for an optimal value of parameter  $\xi_{10}$ . The random component of the error tracking is equal to:

$$\varepsilon^0 = \left[ 1 + \frac{b_1 A(D)}{1 + A(D)B_0(D)} \Phi_1(D) \right] \frac{1}{1 + A(D)B_0(D)} (X^0 + Z^0) \quad (49)$$

where the magnitude  $b_1$  according to formula (36) is given by

$$b_1 = \frac{\partial B_0(\xi_0)}{\partial \xi_{10}} m_{e_0} \quad (50)$$

In computing the dispersion of error  $\varepsilon$  one obtains the formula:

$$D_\varepsilon = \int_{-\infty}^{\infty} \left| \left[ 1 + \frac{b_1 A(i\omega)}{1 + A(i\omega)B(i\omega)} \Phi_1(i\omega) \right] \frac{1}{1 + A(i\omega)B(i\omega)} \right|^2 [S_x(\omega) + S_z(\omega)] d\omega \quad (51)$$

The calculations carried out for a tracking system (Figure 4) having the values of the preceding example for  $\lambda = 10^5$ ,  $T = 1.0$ , and the optimal value of parameter  $\xi_{10} = 3.0$ , show a sufficiently

good effectiveness of tuning. Thus, the mathematical expectation of tuned parameter  $\xi_1$  is equal to  $m_{\xi_1} = \xi_{10}$ , and the dispersion of the error of tuning computed by formula (48) is given by  $D_{\xi_1} = D_{u_1} = 4 \times 10^{-7}$ . From these calculations it follows that the maximum relative error of tuning the parameter  $\xi_1$ , is equal to  $6.3 \times 10^{-2}$  per cent. As regards the tracking error by the follow-up system, the mathematical expectation of this error in tuning coincides with the value of this magnitude in an optimal system  $m_e = m_{e_0} = 0.33 \times 10^{-2}$ .

The dispersion of the tracking error in a self-tuning system computed by formula (51) coincides with a precision to three significant figures with a value of dispersion of the error of tracking in the optimal system  $D_e \approx D_{e_0} = 2.31 \times 10^{-5}$ . Thus, in the considered example the self-tuning system with the utilization of the method of auxiliary operator assures an effective tuning for the minimum of the second initial moment of error in the presence of random disturbances.

### Conclusion

The considered scheme of a self-tuning system may be effectively utilized both for the direct control of plants and the synthesis of automatic control systems during their design. The advantages of the system of self-tuning utilizing the method of auxiliary operator are: relative simplicity of achieving tuning circuits, effectiveness of operation in the presence of disturbances, and the possibility of obtaining high values of quick response.

### References

1. FELDBAUM, A. A. *Computers in Automatic Control Systems*. 1959. MOSCOW; GIFML
2. MARGOLIS, M. and LEONDES, C. T. A parameter tracking servo for control systems. *Trans. Inst. Radio Engrs*, N. Y. AC-4, N 2 (1959)
3. MARGOLIS, M. and LEONDES, C. T. On the theory of adaptive control systems; the learning model approach. *Automatic and Remote Control*. 1961. London; Butterworths
4. PUGACHEV, V. S. Theory of random functions and its application to problems of automatic control. 1960. Moscow; GIFML

### DISCUSSION

P. EYKHOFF, *Technological University, Delft, Netherlands*

Professor Kazakov and Dr. Evlanov presented an interesting contribution to the theory of self-adjusting systems. Basically their approach is an application of Meissinger's parameter-influence coefficients<sup>1, 2</sup> to a type of problem for which, to the best of my knowledge, they have not yet been applied. The auxiliary operators introduced in the paper can be considered as instrumentations of Meissinger's sensitivity equations.

Besides the general questions related to speed of convergence and stability, there are to my mind a few questions that have to do with assumptions made in the paper.

First it is 'assumed that the tracking model carries out its function in an ideal manner'. Two other authors, Margolis and Leondes, employ the same type of approach to parameter tracking. From their work one notices that under favourable conditions and for the simple case of one parameter to be adjusted it still takes an interval of the order of five process-time constants to bring the model parameter close to its desired value. Under certain conditions even stability in

the small is not assured. For this reason I am a bit worried by this assumption of ideal tracking.

In the second place I am concerned about the validity of the assumption stated as 'representing the statistical quality index  $I^*$  of control in the vicinity of the investigated extremum by a quadratic form in terms of deviations  $u_i = \xi_i - \xi_{i0} \dots$ '. No one will question that, mathematically speaking, an ordinary extremum may be approximated by a quadratic form. For the simple parameter tracking problem of Margolis and Leondes, however, one finds the relation of Figure 4 between  $I^*$  (or better  $J^*$ ) and  $(\xi_i - \xi_{i0})/\xi_{i0}$ . This leads to the question whether in an engineering sense the region over which the approximation holds may not become trivially small.

Finally, I would like to pay some attention to the auxiliary operators.

For the sake of simplicity let us take:

$$G_i = \frac{\partial A(\xi)}{\partial \xi_i}$$

An arbitrary example with the parameter  $\xi_i$  in the denominator of the transfer function will show that the operators  $G_i$  may become rather complicated. This is not quite the case if the auxiliary operator is constructed according to Meissinger<sup>1, 2</sup>. In that case the inputs of the auxiliary operators are the dependent variables from the model  $A(\xi)$ , i.e. the model  $A(\xi)$  is a part of the auxiliary operator. A further simplification is obtained when instead of  $A(\xi)$  a 'generalized model' according to Figure B is allowed<sup>3</sup>. Then there is no need for separate auxiliary models and one arrives at the interesting model-adjustment

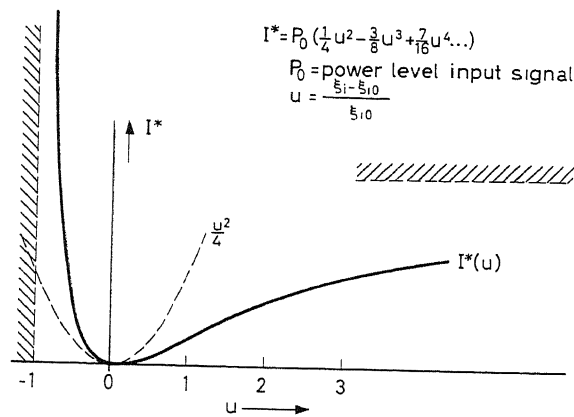


Figure A

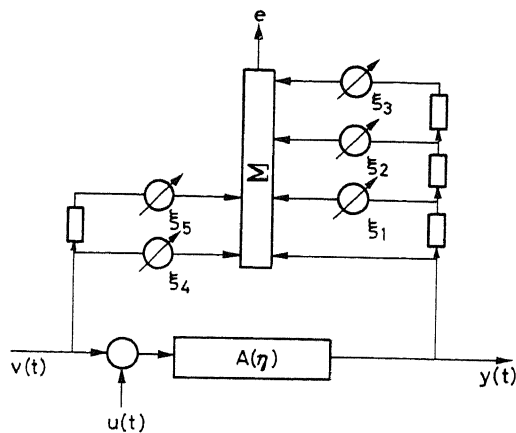


Figure B

schemes of Clymer<sup>4, 5</sup> and Graupe<sup>6</sup>. Under some conditions even a fast, monotonic convergence of the error  $e$  to zero can be assured<sup>6</sup> irrespective of the type of process-input signal. Additive noise, however, leads to a bias in the adjusted parameters. Means to eliminate this drawback will be reported in the near future.

## References

- <sup>1</sup> MEISSINGER, H. F. The use of parameter influence coefficients in computer analysis of dynamic systems. *Proc. West. Joint Computer Conf.*, San Francisco (May 1960) 181-192
- <sup>2</sup> MEISSINGER, H. F. Parameter influence coefficients and weighting functions applied to perturbation analysis of dynamic systems. *Third Int. Congr. Analogue Computation, Opatija, Yugoslavia* (1961)
- <sup>3</sup> EYKHOFF, P. Some fundamental aspects of process-parameter estimation. *Rep. Electron. Lab., Tech. Univ. Delft.* (Dec. 1962); *I.E.E.E. Trans. Automatic Control* (Oct. 1963)
- <sup>4</sup> CLYMER, A. B. Direct system synthesis by means of computers. *Communication and Electronics; Trans. Amer. Inst. Elect. Engrs.* Pt. I, 77 (1959) 798-806
- <sup>5</sup> POTTS, T. F., ORNSTEIN, G. N. and CLYMER, A. B. The automatic

determination of human and other system parameters. *Proc. Joint Computer Conf.*, Los Angeles (May 1961) 645-660

- <sup>6</sup> GRAUPE, K. K. The analogue solution of some fundamental analysis problems. *Communication and Electronics; Trans. Amer. Inst. Elect. Engrs.* Pt. I, 52 (1961) 793-799

J. ZABORSKY and D. GORMAN, *Washington University, St. Louis 30, Missouri, U.S.A.*

(1) The paper consistently uses the terms 'operator' and 'functional' without stating the limitations imposed on these operators by the manipulations used in the paper. The most striking example is that  $N$  and  $L$  are simply identified as 'an operator or a functional', which enter into the indices of control and appear to be restricted linear, instantaneous operators. For a deterministic system  $N$  and  $L$  may seem limited to scalar constants; for a random system  $N$  and  $L$  may be expectation operators. Time integrals such as integral square criteria are not apparently permitted, however, since these change the derivatives of  $I$  and  $J$  into functional derivatives, thus altering the nature of the problem. All the other operators, especially  $A(\eta)$  and  $B(\xi)$ , seem to be limited to be linear and dependent only on instantaneous values of the arguments.

(2) What is the validity of the stability analysis in the paper considering that  $A(\eta)$  was substituted for  $A(\xi)$ . This could introduce delays and inaccuracies which may strongly affect stability.

(3) With regard to the example, what would make it necessary to use the second initial moment as a performance criterion?

(4) Once the assumption is made in the example that noise  $Z$  can be essentially separated from useful signal  $X$  it would appear logical to use this filtered  $X$  signal directly for control and avoid the complex secondary structure involving  $Z_1$ , as introduced in the paper and shown in Figure 4. This difficulty does not seem limited to this specific example but appears general for the method since only signal  $e$  would be normally available for direct measurement with the assumptions of the paper.

J. D. ROBERTS, *University of Cambridge, Cambridge, England*

The method of the auxiliary operator will, I believe, have important applications in designing self-adjusting control systems. A similar idea is used in the sensitivity function used by Olsum in modelreference adaptive systems. I also mention my paper, published in the *I.E.E.* in 1962, which embraces this kind of system. My first point is that the auxiliary operator is realized by a network (called the 'auxiliary network') which is related to the original network in a precise way. Consider the example given by the authors:

An auxiliary network is easily constructed in which signals corresponding to  $E$  and  $V$  in the original network have values  $\partial e / \partial \xi_1$  and  $\partial V / \partial \xi_1$ . The loop is duplicated and a connecting link is provided  $\partial B / \partial \xi_1 = 1$  realized by a straight connection. Non-linear characteristics can also be treated. In particular, if a saturation characteristic is introduced after the signal  $V$ , a switch must be introduced after the signal  $\partial V / \partial \xi_1$ . This must be in the break position at times when the non-linear element is saturating. As a consequence of a discussion with Professor Pugachev, I point out that the signal  $\partial e / \partial \xi_1$  can also be obtained by a connecting link which adds the signal  $V$  to the input of the auxiliary  $B(\xi)$  (the dotted connection) instead of adding the signal  $e$  to the output of the auxiliary  $B(\xi)$ . However, this alternative fails when non-linearities are introduced.

Finally, a point concerning the required accuracy of the auxiliary  $A(\eta)$ . In this respect, the method is especially powerful with cascade or open-loop controllers, because the accuracy is not critical. As an extreme case, consider the open-loop control of a system with known dynamics  $A$ , but unknown gain  $\eta_1$ . The auxiliary network is shown with an inaccurate gain  $\eta_1^*$ . The criterion of zero mean product of the signals is unaffected by an error in  $\eta_1^*$ . More specifically, the value of  $\xi_1$  satisfying this criterion will be intermediate between the time optimum and the optimum for the inaccurate  $\eta_1^*$ .

P. KOKOTOVIĆ, *Institute for Automation and Telecommunication, Belgrade, P.O.B. 906, Yugoslavia*

Taken generally, the two basic problems in the synthesis of self-adapting systems, namely process identification and parameter adjustment, can be solved using gradients of the properly chosen index of quality  $I$  and index of approximation  $J$ . Applicability of such an approach depends to a great extent on the effectiveness of the method by which the components of the vectors grad  $I$  and grad  $J$  are determined during the operation of the system. In this regard the parameter influence technique offers promising possibilities<sup>1-10</sup>.

The paper by Kazakov and Evlanov represents an excellent theoretical generalization of the application of the parameter influence technique to the synthesis of self-adapting systems<sup>11, 12</sup>. It should be pointed out, however, that the structure shown in Figure 2 of the paper increases in complexity as the number of the parameters to be adjusted increases. We would, therefore, suggest a procedure<sup>8, 9</sup> which avoids this difficulty, at least as far as linear systems are concerned, and enables all parameter influences to be computed simultaneously by means of one single structure that could be called the parameter influence analyser.

#### The Structure of the Parameter Influence Analyser

Let the system transfer function  $W = W(s, q)$  depend on parameters  $q_i$ ,  $i = 1, \dots, 2m$ . Let  $W_i = W_i(s, q_i)$  be the transfer function of the  $i$ th system element containing the parameter  $q_i$ . Then the system equation

$$Y(s, q) = X(s) W(s, q) \quad (1)$$

can be differentiated as follows

$$\frac{\partial Y}{\partial \ln q_i} = Y \frac{\partial \ln W}{\partial \ln W_i} \frac{\partial \ln W_i}{\partial \ln q_i} \quad (2)$$

where

$$Z^{-1} \frac{\partial Y}{\partial \ln q_i} = Z^{-1} U_i(s) = u_i(t) \quad (3)$$

are the influences of the relative changes of parameters on the system response  $y(t, q)$ . According to eqns (2) and (3) the influence  $u_i(t)$  is computed from the response  $Y(s, q)$  and the transfer functions

$$P_i = \frac{\ln W}{\ln W_i} \quad \text{and} \quad F_i = \frac{\ln W_i}{\ln q_i} \quad (4)$$

It can be shown that the transfer function  $P_i$  determines the position of the point  $S_i$  in the structure of the system  $W(s, q)$ . We shall call these the sensitivity points. It follows that all parameter influences  $u_i(t)$ ,  $i = 1, \dots, 2m$ , can be obtained simultaneously by means of a single structure that consists of one main part identical with the system structure and several additional parts whose transfer functions are  $F_i$ , and which are connected to the main part at the corresponding sensitivity points  $S_i$ .

To illustrate this procedure we use the system shown in Figure A. Positions of sensitivity points in this general multi-loop feedback structure are indicated in Figure B. The parameter influences are obtained at the outputs of  $F_i$ .

It should be emphasized that  $F_i$  are usually of a simple structure. In the most important practical case, i.e. when the parameter  $q_i$  is the gain factor of the element  $W_i$ ,  $F_i$  becomes equal to unity. Hence, the parameter influences are obtained directly at the sensitivity points and the structure of the parameter influence analyser becomes identical with the structure of the system.

It follows, therefore, that in many practical cases blocks  $C_1, \dots, C_n$  in Figure 4 of the paper can be substituted by a single structure which usually is no more complex than the structure of the system itself.

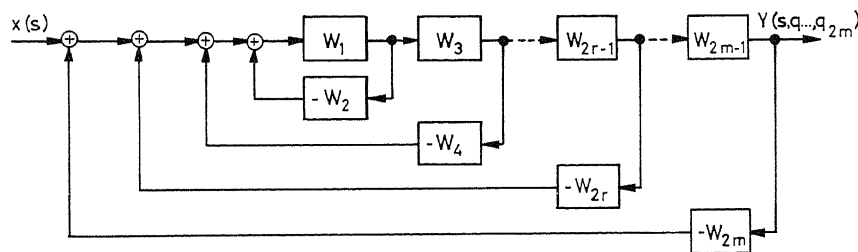


Figure A

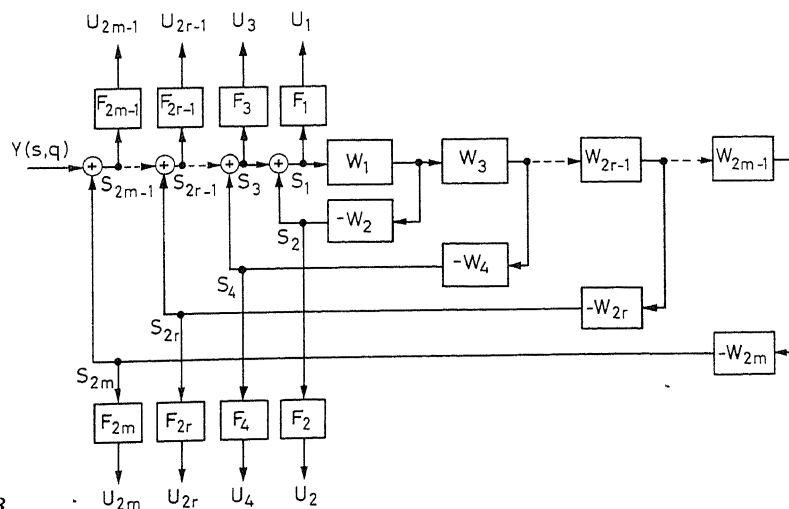


Figure B

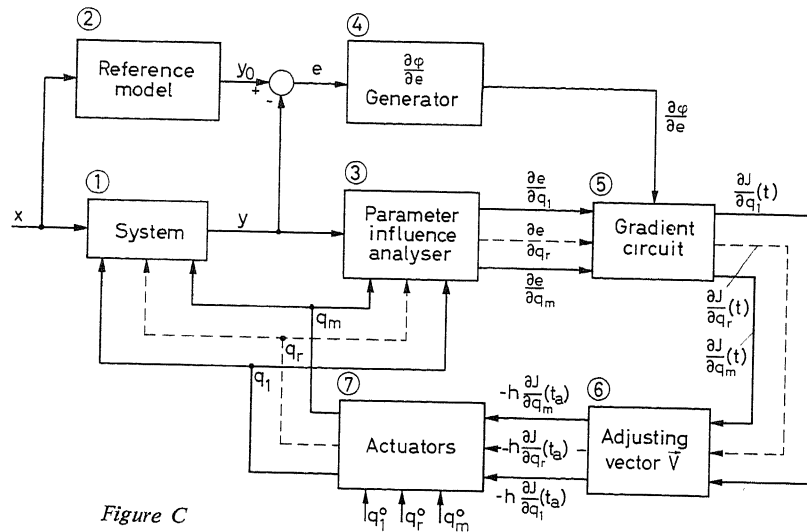


Figure C

The similar case applies to blocks  $G_1, \dots, G_n$ . In this way the realization of an adaptive system with multiple adjustable parameters can be greatly simplified<sup>13</sup>.

#### Application to Automatic Design

Such a procedure has been applied also to the automatic design of control systems using a general purpose analogue computer<sup>14</sup>. Basic diagram of the analogue computer circuit is shown in Figure C. Operating modes of the blocks 1, ..., 5 are automatically switched in the sequence 'operate-hold-init. condition', while the operating sequence of blocks 6 and 7 is 'hold-operate-hold'. The computing circuit can be still further simplified if a repetitive computer equipped

with servomultipliers is used. Figure D is an example of optimization process for a system having a structure as in Figure A. Parameter values are rapidly found by the computer, even in cases where the starting point falls in an unstable region.

Our experience in solving similar problems confirms the opinion of the authors that the method described can be successfully used for automatic design of control systems.

#### References

- VASILJEVA, A. B. On the differentiability of the solution of differential equations containing small parameters (in Russian). *Dokl. Akad. Nauk S.S.S.R.* 61 (1948) 597-601
- MILLER, K. S. and MURRAY, F. J. A mathematical basis for an error analysis of differential analysers. *J. Math. Phys.* 32
- BIHOYSKI, M. L. *Principles of Dynamic Accuracy of Electrical and Mechanical Networks*. 1958. Moscow; Academy of Sciences
- MEISSINGER, H. The use of parameter influence coefficients in computer analysis of dynamic systems. *Proc. West. Joint Computer Conf.* San Francisco, May 1960
- MEISSINGER, H. Parameter influence coefficients and weighting functions applied to perturbation analysis of dynamic systems. *Third Int. Conf. Analog Computation, Opatija*, 1961
- CHANG, S. S. L. *Synthesis of Optimum Systems*. Chap. 8. 1961. New York; McGraw-Hill
- TOMOVIĆ, R. and KARPLUS, W. *High Speed Analog Computers*. Chap. 4. 1962. New York; Wiley
- KOKOTOVIĆ, P. Parameter influence analysis by structural method. *Magistar Thesis*, Fac. of Elect. Engng, Univ. Belgrade, 1962 (in Serbian)
- KOKOTOVIĆ, P. Parameter influence analysis of linear feedback systems. *8th Yugoslav Conf. E.T.A.N. Zagreb*, 1963
- TOMOVIĆ, R. *Sensitivity Analysis of Dynamic Systems*. McGraw-Hill (in press)
- MARGOLIS, M. and LEONDES, C. T. A parameter tracking servo for control systems. *Trans. Inst. Radio Engrs. N.Y. AC-4*, N 2 (1959)
- MARGOLIS, M. and LEONDES, C. T. On the theory of adaptive control systems; the learning model approach. *Automatic and Remote Control*. 1961. London; Butterworths
- BINGULAC, S. General procedure for synthesis of multi-parameter adaptive control systems. *Inter. Rapp. Princeton Computing Center* (1962)
- BINGULAC, S. and KOKOTOVIĆ, P. Automatic optimization of control systems using general purpose analog computer. *8th Yugoslav Conf. E.T.A.N. Zagreb*, 1963

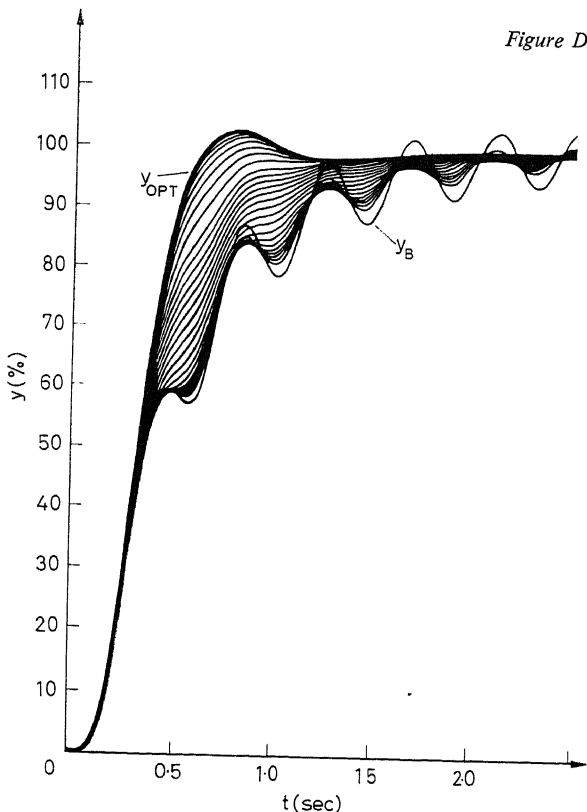


Figure D

# Problems of Continuous Systems Theory of Extremal Control of Industrial Processes

A. A. KRASOVSKI

## Summary

An extremal control of a continuous industrial process is discussed. A control system maintains the extremum of some index of the product quality. Equations of continuous extremum control systems are analysed. The effects of delay and disturbances on the control processes are determined.

Effective circuits for the production parameter control are discussed. Possibility and expediency of introducing parametric self-adjustment of extremum control systems for complex industrial processes are shown.

## Sommaire

On examine le problème de commande automatique pour obtenir l'extrémum d'un indice de qualité d'un processus industriel continu. On analyse les équations générales d'extrémalisation les effets des retards et des perturbations. On examine les possibilités et l'efficacité de certains schémas de commande automatique extrémalisante pour des processus industriels complexes.

## Zusammenfassung

Die Extremwertregelung von Maschinen, Geräten und Reglereinstellungen, die gemeinsam einen kontinuierlichen industriellen Prozeß bilden, wird besprochen. Eine Regelung sorgt für die Einhaltung des Extrems eines Indexes der Produktqualität. Es werden Gleichungen für kontinuierliche Extremwertregelungen untersucht. Die Auswirkungen

von Verzögerungen und Störungen auf die Regelung werden ermittelt.

Brauchbare Regelkreise für die Fertigungskennwerte werden besprochen. Die Möglichkeit und die Brauchbarkeit der Selbsteinstellung der Parameter bei Extremwertregelungen wird für komplexe Industrieprozesse gezeigt.

Many continuous industrial processes lend themselves to the following plan. There is available some quantity  $n$  of adjustments or controls of machines, apparatus, regulators securing an industrial process. The flow of the industrial process and the output parameters depend on the coordinates of the adjusting or control elements (adjustment parameters).

Together with the control adjusting element coordinates the output parameters are affected by various disturbance factors (change of material parameters, wear of machines and tools, temperature and moisture variations and other factors).

The output parameters are controlled continuously or discretely (but with sufficiently small intervals of discontinuity) by special measuring devices—output parameter information transmitters (Figure 1). Influenced by disturbance factors and also by random variations of adjustment parameters the output parameters are subjected to continuous variations.

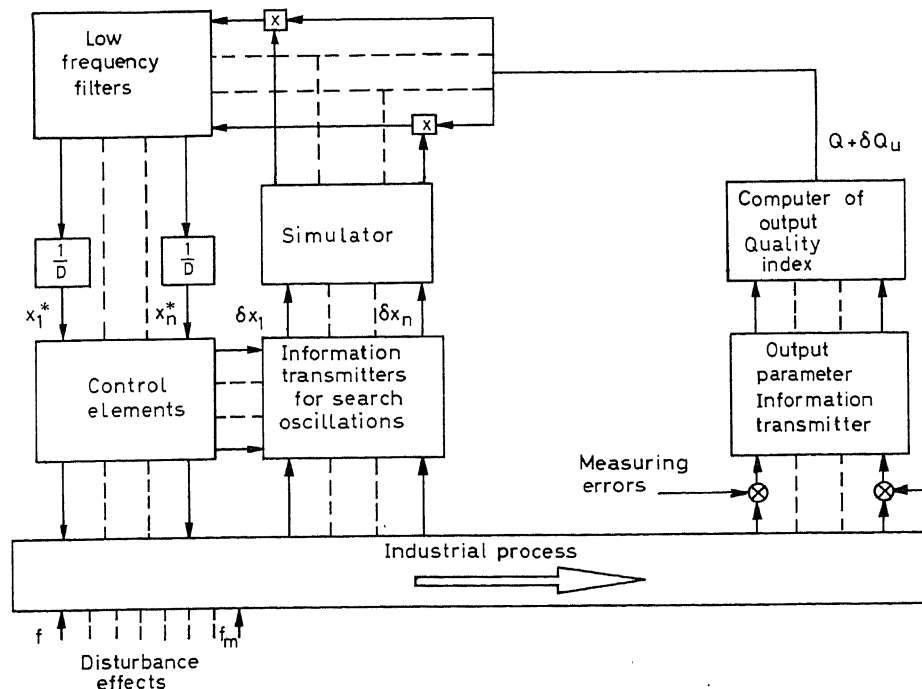


Figure 1. Layout of an extremal control system of an industrial process with a single criterion of production quality

Even though a practically ideal adjustment of the producing system, ensuring the industrial process, is initially attained, after some time the disturbance factors will bring forth considerable changes in the output parameters. In order to prevent the dropping out of the output parameters from the established tolerances (scrap output), adjustment and tuning of the production system is necessary. Various means of automation of these operations are possible. If it is precisely known which parameter and to what extent is affected by one or another controlling element, the usual feedback principle may be used (control by deviation). For this it is necessary first to smooth the results of measurements in order to eliminate overshoots of the system in the presence of small, random deviations within the tolerance limits. Methods of such automatic processing of information may be set up, based on the widely utilized methods of non-automated statistical control<sup>2</sup>. The measured and smoothed signals of output parameter deviations are conveyed to the performing arrangements and cause changes in the controlling element coordinates. Such systems are sometimes called statistical automata<sup>3</sup>.

Undoubtedly the introduction of statistical automata will prove to be an important step in the automation of industry. However, a necessary condition of their application must be a sufficiently complete *a priori* information on the characteristics of the industrial process. In many cases this information is absent, and even if it is available during the initial period of the systems adjustment it loses authenticity in time, due to the change in properties of the industrial process.

Under these conditions the application of usual, non-self-adjusting control loops (statistical automata) becomes impossible. In these cases it is expedient to utilize an extremal control.

The present work is devoted to the investigation of some possible circuits of extremal control systems with continuous industrial processes and some questions of the theory of these systems. It is a development of earlier work by the author<sup>1</sup>.

For the realization of an extremal control a quality output (production) index  $Q$  is selected, having extrema at wanted values of product parameters. Such an index may be, for example, the sum of the squares of deviations of the output parameters from the standard values. The quality index  $Q$  is determined by a computer in diagram *Figure 1*) based on information transmitter data on current values of output parameters. To secure the basic function of the system-maintenance of the quality index at the extremum level, search oscillations are necessary. Natural high frequency random oscillations, as well as artificially produced oscillations of controlling elements, may be employed as search oscillations. Naturally the first method is preferable, since it is not linked with any increase of high frequency fluctuations of the production parameters.

In order to make use of natural oscillations as search oscillations, it is necessary to measure them. The measurement of search oscillations is done by information transmitters for these oscillations (*Figure 1*), which measure controlling element oscillations and disturbance effects transmitted to them.

The measured search oscillations are transmitted to a simulator or a dynamic model of the industrial process. The purpose of the simulator is to transform the search oscillations in the same manner as these oscillations are transformed in a real process. For many industrial processes the simulator may be carried out in the shape of a delay line.

The output signals of the simulator are transmitted to the multiplying elements, to the other entrances of which is transmitted the computer signal which is proportional to the current value of the production quality index. The output values of the multiplying element are smoothed by the low frequency filters and are transmitted to the entrances of the control devices which move the controlling element.

If the quality index deviates from the extremum value, then a component correlated with the search fluctuations appears at the computer output. Values of the mathematical expectation of the duplicating links signals differing from zero then appear. Slowly changing signals are separated out by the low frequency filters and start the control devices. The controlling elements act on the production parameters in the direction approaching the extremum of the quality index.

The values of control parameters, together with the disturbance effects transmitted to them, are designated as  $X_v$  ( $v = 1, 2, \dots, n$ ). Each control parameter brought forth has three components.

$$X_v = X_v^* + \delta X_v + \delta X_{vw}$$

Here  $X_v^*$ , working elements, are output values of the extremal control part of the system;  $\delta X_v$  are search elements for which it is expedient to utilize high frequency controlled effects transmitted to the control parameters, and  $\delta X_{vw}$  are uncontrolled disturbance effects transmitted to the control parameters.

The current value of the production quality index in general is a function of indicated control parameters and disturbance effects  $f_1, f_2, \dots, f_m$  according to transmitted control parameters.

When the transient process characteristics are described with sufficient accuracy by time delays, then the current value of the production index is expressed by the function of preceding values of indicated control parameters and disturbances effects.

$$Q = Q[X_1(t - \tau_1), \dots, X_n(t - \tau_n), f_1, \dots, f_m] \quad (1)$$

The selection of the composition of control parameters must conform to the following condition. To each set of permanent control parameter values must correspond a definite (with an accuracy up to the level of noises) set of production parameter values. In other words, in a static regime and with absence of noises a simple conversion of control parameters into production parameters must be realized. It should be noted that no mutual unilateral conversion is required, so that the number of control parameters may greatly exceed the number of controlled production parameters.

In virtue of one-sided-unilateral conversion, to each extremal function of production parameters, corresponds an extremal function of control parameters.

As agreed, the production quality index is an extremal function of its parameters. Therefore, function (1) in relation to the control parameters  $X_1, X_2, \dots, X_n$  is also extremal.

#### Adjustment-loss Time

Assuming that a process having unchanged, fixed working components of control parameters  $X_v^*$ , is under investigation, and assuming also that, by the initial adjustment, it was possible, at some time  $t = t_0$ , to attain the extremum value of



the production quality index, then under the influence of disturbance factors the production quality index will in time deteriorate spontaneously, in spite of the constancy of the control coordinates (Figure 2). At the expiration of time  $T_1$  the quality index will get out of the permissible limits. The disturbance effects are random functions of time or random values, although in some individual applications their mathematical expectations may dominate centred random elements.

The change of production quality index with time  $Q(t)$  is also a random time function, known to be non-stationary for this process with a fixed adjustment. And so, repeating the above test, one gets new realizations  $Q(t)$  and new time values  $T_i$  (Figure 2).

The overall adjustment-loss time by the quality index  $Q$  is designated as the mathematical expectation  $M(T_i)$  of time intervals  $T_i$ . So the overall adjustment-loss time expresses the mean value of time interval, after which the production quality index of the industrial process with a fixed adjustment gets out of the permissible limits.

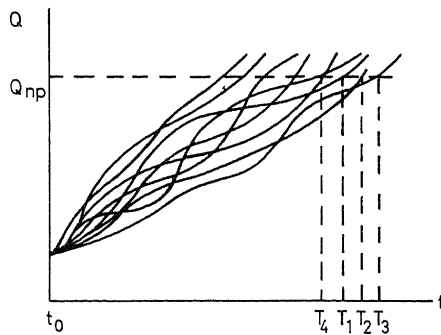


Figure 2. For determining the conception of general disadjustment time

The adjustment-loss time, understandably, depends on the nature of the industrial process and its automation level by means of frequency automatic systems. If the overall adjustment-loss time is large, then a non-automatic, hand control is not difficult and there is no need to use a complex self-adjusting system. If the overall adjustment-loss time is small, then a person is unable to secure adjustment even with the presence of appropriate data transmitters and self-adjustment becomes necessary.

It should be noted that the higher the speed of the industrial process and the stricter the demands on the quality of production, the smaller is the overall adjustment-loss time. Acceleration of the industrial processes and stepping up of demands on the quality of production are inherent characteristics of technical progress. Therefore, the application of self-adjusting control systems to industrial processes has a broad prospect.

### Equations of Extremal Control Processes

It is assumed that, in the vicinity of the extremum point, serving as a work region of the system under consideration, the quality index (1) approximates with sufficient accuracy by the quadratic form of preceding values of control parameters and by the additional member  $\delta Q_f$  expressing the influence of disturbing effects  $f_1, \dots, f_m$ :

$$Q(t) = Q_e + \frac{1}{2} \sum_{i,j=1}^n a_{ij} \Delta X_i(t-\tau_i) \Delta X_j(t-\tau_j) + \delta Q_f$$

$$a_{ij} = a_{ji} = \frac{\partial^2 Q}{\partial X_i \partial X_j} \quad (2)$$

here

$$\Delta X_v = X_v - X_{vl} = X_v^* + \delta X_v + \delta X_{vw}$$

$$-X_{vl} = \Delta X_v + \delta X_v + \delta X_{vw}$$

are complete deviations of brought forth control coordinates (parameters),  $\Delta X_v^* = X_v^* - X_{vl}$  are working deviations of control coordinates, and  $Q_e$  is the extremum value of the quality index. In case the computer of the quality index does not ensure smoothing (which is secured only by subsequent elements of the circuits) and the production parameter measuring instruments are practically non-inertial, or their inertness is accounted for in the values of time delays  $\tau_i$ , the output value of the computer equals:

$$U(t) = Q(t) + \delta Q_n(t)$$

Here  $\delta Q_n(t)$  is the element conditioned by the errors of the production parameter meters and the errors of the computer.

Thus

$$U(t) = Q_t + \frac{1}{2} \sum_{i,j=1}^n a_{ij} \Delta X_i(t-\tau_i) \Delta X_j(t-\tau_j) + \delta Q \quad (3)$$

where

$$\delta Q = \delta Q_f + \delta Q_n$$

The value  $U(t)$  in the multiplying elements of the synchronous detectors (correlators) is multiplied by the search signals  $\delta X_v$ , displaced in time by the delay simulator. The errors in delay simulation are designated  $\delta \tau_v$ .

To the second entrances of the multiplying elements are transmitted values  $\delta X_1(t-\tau_v-\delta \tau_v)$  and the output signals of these elements equal  $V_v = U(t) \delta X_v(t-\tau_v-\delta \tau_v)$ .

The linear portion of the controlling system without any common restriction is divided into a set of filters and integrating elements (Figure 1). The output working coordinates equal

$$X_k^* = \frac{1}{D} \sum_{v=1}^n W_{kv}(D) V_v$$

Here  $\|W_{kv}(D)\|$  is the matrix of the transfer functions at low frequencies. Thus

$$DX_k^* = D\Delta X_k^* + DX_{kl} = \sum_{v=1}^n W_{kv}(D) V_v, \quad D = \frac{d}{dt}$$

or

$$D\Delta X_k^* = \sum_{v=1}^n W_{kv}(D) [u(t) \delta X_v(t-\tau_v-\delta \tau_v)] - DX_{kl}$$

utilizing equations (3) for  $U(t)$ , one finds

$$D\Delta X_k^* = \frac{1}{2} \sum_{i,j,v} a_{ij} W_{kv}(D) \{ [\Delta X_i^*(t-\tau_i) + \delta X_i(t-\tau_i) + \delta X_{iv}(t-\tau_i)] \times [\Delta X_j^*(t-\tau_j) + \delta X_j(t-\tau_j) + \delta X_{jv}(t-\tau_j)] \times \delta X_1(t-\tau_v-\delta \tau_v) \}$$

$$+ \sum_v W_{kv}(D) [(Q_t + \delta Q) \delta X_v(t-\tau_v-\delta \tau_v)] - DX_{kl} \quad (4)$$

$$(k=1, 2, \dots, n)$$

Summation by indices  $i, j, v$  is carried out within the limits from 1 to  $n$ .

## Qualitative Analysis of Extremal Control Processes

### Quasi-stationary Regime

The quality demand of an extremal control process reduces to the following. With considerable initial deviations from extremum the state point must move to the extremum as smoothly as possible (without much overshoot). In a steady operation the state point must stay sufficiently close to the extremum.

Let eqn (4) be converted into:

$$\begin{aligned}
 D\Delta X_k^* &= \sum_{i,j,v} a_{ij} W_{kv}(D) \\
 &\quad [\Delta X_i^*(t-\tau_i) \delta X_j(t-\tau_j) \delta X_v(t-\tau_v-\delta\tau_v)] \\
 &+ \frac{1}{2} \sum_v W_{kv}(D) \\
 &\quad \sum_{i,j} a_{ij} \Delta X_i^*(t-\tau_i) \Delta X_j^*(t-\tau_j) \delta X_v(t-\tau_v-\delta\tau_v) \\
 &+ \sum_{i,j,v} a_{ij} W_{kv}(D) \\
 &\quad [\Delta X_i^*(t-\tau_i) \delta X_{jw}(t-\tau_j) \delta X_v(t-\tau_v-\delta\tau_v)] \\
 &\quad + \delta f_k - D X_{ki}
 \end{aligned} \quad (5)$$

here

$$\begin{aligned}
 \delta f_k &= \frac{1}{2} \sum_{i,j,v} a_{ij} W_{kv}(D) \{ [\delta X_{iw}(t-\tau_i) + \delta X_i(t-\tau_i)] \\
 &\quad \times [\delta X_{jw}(t-\tau_j) + \delta X_j(t-\tau_j)] \delta X_v(t-\tau_v-\delta\tau_v) \} \\
 &+ \sum_v W_{kv}(D) [(Q_c + \delta Q) \delta X_v(t-\tau_v-\delta\tau_v)]
 \end{aligned} \quad (6)$$

Values  $\delta f_k$  may be treated as the effect of errors, noises and search elements, brought to the outputs of filters of the synchronous detectors, provided there are no working deviations ( $\Delta X_i^* = 0$ ). These functions do not depend on working deviations (it is assumed, that  $\delta Q$  does not depend on working deviations) and on the whole may only hinder the movement of the state point to the extremum.

Thus,  $\delta f_k$  always play the role of disturbance effects and it is expedient to decrease them as much as possible. If the search elements  $\delta X_j$ , have permanent constituents then, as seen from expression (6), it is impossible to decrease indefinitely  $\delta f_k$  by any increase of time constant filters of the synchronous detectors. Indeed, according to (6), the constant components  $\delta X_v$  will cause deviations at the outputs of the synchronous detectors

$$\frac{1}{2} \sum_{i,j,v} a_{ij} W_{kv}(0) \overline{\delta X_i} \overline{\delta X_j} \overline{\delta X_v} + \sum_v W_{kv}(0) (Q_{kv} + \overline{\delta Q}) \delta X_v$$

where at least some of the transfer coefficients  $W_{kv}(0)$  are known to differ from zero, since otherwise the extremal control circuit inefficient. Thus, it is expedient to secure zero parity of the permanent elements of search constituents i.e. the centring of the search oscillations. This is easily attained by installation of high frequency filters at the outputs of the search oscillation pickups.

In particular, an ideal high frequency filter separates, from the input value, the high frequency constituent uncorrelated

with the remaining part of the input value. This is illustrated by the graphs in Figure 3, showing a density spectrum curve  $S(\omega)$  of the input function, which is assumed to be stationary and ergodic and an amplitude frequency characteristic  $A(\omega)$  of an ideal high frequency filter.

An ideal filter separates the high frequency constituent with a spectral density  $S\delta_v(\omega)$  Figure 3(b) uncorrelated with the filtered component (spectral density  $S_w(\omega)$ ), since the mutual spectral density of these components equals zero.

If the data meter controls the full input coordinate of the system  $X_v = X_v^* + \delta X_v + \delta X_{vw}$ , then the ideal high frequency filter in a stabilized operation separates the high frequency constituent  $\delta X_v$  uncorrelated with the constituent  $X_v^* + \delta X_{vw}$ . It should be noted that stationary  $X_v^*$  may be expected only in a stabilized regime of the system operation. In transient regimes  $X_v^*$  is a non-stationary random function and even with the use of ideal filters the search elements prove to be to some extent correlated with the working elements  $X_v^*$ .

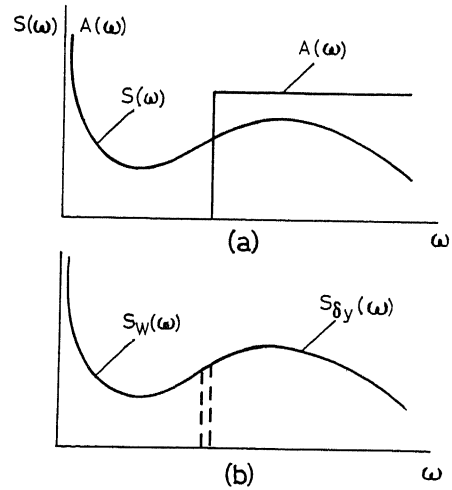


Figure 3. Separation of an independent search component by an ideal high frequency filter

However, as is seen from the following in the present system (perhaps even more than in other continuous extremal systems), a quasi-stationary regime is profitable. In a quasi-stationary regime the transient process times are large compared to correlated times of search elements. When a quasi-stationary condition is secured and with the application of high frequency filters near to ideal the search elements may be considered with a high degree of accuracy as uncorrelated with  $X_v^*$ , both in a stabilized and in a transient condition of the system.

Based on the above the search elements  $\delta X_v$  is assumed to be centred by random functions as uncorrelated to  $\Delta X^*$ ,  $\delta X_{vw}$ ,  $\delta Q$ . Investigation of other members of the right portions of eqn (5) is now made. The second member of the right portion may be rewritten in the shape

$$\begin{aligned}
 &\frac{1}{2} \sum_v W_{kv}(D) [F^* \delta X_v(t-\tau_v-\delta\tau_v)] \\
 &\text{where} \\
 &F^* = \sum_{i,j} \Delta X_i^*(t-\tau_i) \Delta X_j^*(t-\tau_j)
 \end{aligned} \quad (7)$$

In view of the definiteness of the signs of the functions of working deviations this member cannot facilitate the organization of movement to the extremum.

Thus, members (7) play the part of impeding effects and it is expedient to reduce their influence to minimum values.

The only accepted means of reducing the effects of these members is the raising of frequencies (decreasing correlation times) of the search elements at given times of transient processes of a closed loop or, inversely, increasing cumulative times at given correlation times of search elements. Either one or the other means result in a quasi-stationary regime. In a quasi-stationary regime the effects of members (7) can be neglected. The following members of eqns (5)

$$\sum_{i,j,v} a_{ij} W_{kv}(D) [\Delta X_i^*(t-\tau_i) \delta X_{jw}(t-\tau_j) \delta X_v(t-\tau_v-\delta\tau_v)] \quad (8)$$

although linearly depend on working deviations, are also playing the role of impeding effects.

In fact, as agreed  $\delta X_{jw}$  and  $\delta X_v$  are uncorrelated and  $\delta X_v$  are centred. Therefore, the mathematical expectations of the products  $\delta X_{jw}(t-\tau_j) \delta X_v(t-\tau_v-\delta\tau_v)$  equal zero. Thus, the expressions in the square brackets represent linear forms of working deviations, whose coefficients are centred 'high frequency' random time functions. These members can only increase the scattering of the trajectories of the state point during its movement to the extremum.

In the quasi-stationary regime, because of the intensive suppression of the high frequency constituents, members (8) may be neglected.

Turning to the investigation of the first member of the right portions of eqns (5) it is noticed that the product of search constituents may be represented as a sum of the mutual correlated (at  $j \neq v$ ) or an auto-correlated function and a centred random function. Moreover, if the search constituents are stationary and are stationarily combined, then the correlation functions depend only on the argument difference.

$$\delta X_j(t-\tau_j) \delta X_v(t-\tau_v-\delta\tau_v) = R_{jv}(\tau_v-\tau_j+\delta\tau_v) + \xi_{jv}(t)$$

where  $\xi_{jv}(t)$  are centred random function members

$$\sum_{i,j,v} a_{ij} W_{kv}(D) \Delta X_i \xi_{jv}(t) \quad (9)$$

play the same kind of negative role as members (8). In a quasi-stationary regime the influence of these members may be decreased to the same extent, as the influence of members (8), since the correlation times of function  $\xi_{jv}(t)$  are compared with the correlation times of function  $\delta X_{jw} \delta X_v$ . In a quasi-stationary regime one neglects the influence of members (9).

Therefore, the general equations (5) are replaced by the following equations of a quasi-stationary regime of the system under consideration.

$$\begin{aligned} D\Delta X_k^* &= \sum_{i,j,v} a_{ij} R_{jv}(\tau_v-\tau_j+\delta\tau_v) W_{kv}(D) \Delta X_i^*(t-\tau_i) \\ &+ \delta f_k(t) - DX_{kl} \\ (k=1, 2, \dots, n) \end{aligned} \quad (10)$$

These general equations of a quasi-stationary regime are simplified in concrete, particular cases.

First of all it is noted that the correlation times of search signals are small, due to the presence of high frequency filters.

Therefore, for a typical case, when the delay times  $\tau_j$  are not identical it may be assumed

$$R_{jv}(\tau_v-\tau_j+\delta\tau_v) = \begin{cases} 0 & \text{for } j \neq v \\ R_v(\delta\tau_v) & \text{for } j = v \end{cases} \quad (11)$$

It should be noted that the same correlations take place also at strictly identical delay times  $\tau_v = \tau_j$ , but with uncorrelated search constituents. In practice, uncorrelated natural search constituents may be obtained by means of installing instead of high frequency filters, band filters with non-overlapping passing bands. The shortcoming of this method is the considerable lowering of the level or efficiency of the utilized search elements, especially in multi-instrument systems ( $n$  dimensional).

Under condition (11) eqns (10) take the shape

$$D\Delta X_k^* + \sum_{i=1}^n W_{ki}^c(D) \Delta X_i^*(t-\tau_i) = \delta f_k - \dot{X}_{kl} \quad (12)$$

where

$$W_{ki}^c(b) = \sum_{v=1}^n a_{iv} R_v(\delta\tau_v) W_{kv}(D)$$

It is also possible to introduce transfer functions of a closed-loop system, then

$$\Delta X_k^* = \frac{1}{\Delta(D)} \sum_{v=1}^n (-1)^{k+v} \Delta_{kv}(D) (\delta f_v - \dot{X}_{vl}) \quad (13)$$

where

$$\Delta(D) = \begin{vmatrix} D + W_{11}^c(D) e^{-\tau_{1b}} & \dots & W_{1n}^c e^{-\tau_{nb}} \\ \dots & \dots & \dots \\ W_{n1}^c(b) e^{-\tau_{1b}} & \dots & D + W_{nn}^c(D) e^{-\tau_{nb}} \end{vmatrix}$$

$\Delta_{kv}(D)$  is the determinant, obtained from  $\Delta(D)$  by cancelling 'K' column, 'v' line.

The roots of a characteristic equation

$$e^{-(\tau_1 + \dots + \tau_n)\lambda} \begin{vmatrix} \lambda e^{\tau_{1\lambda}} + W_{11}^c(\lambda) & \dots & W_{1n}^c(\lambda) \\ \dots & \dots & \dots \\ W_{n1}^c & \dots & \lambda e^{\tau_{n\lambda}} + W_{nn}^c(\lambda) \end{vmatrix} = 0 \quad (14)$$

are simply determined in case when times  $\tau_v$  are practically equal, the synchronous detector filters are neutral and possess identical transfer functions:

$$R_v(\delta\tau_v) W_{kv}(D) = \begin{cases} 0 & \text{for } v \neq k \\ W(D) & \text{for } v = k \end{cases}$$

In this case the characteristic equation (14) breaks up into  $n$  equations

$$\frac{\lambda e^{\tau\lambda}}{W(\lambda)} = -\frac{1}{C_v^2} \quad (15)$$

where  $C_v$  is the semi-axis of the determining ellipsoid

$$\sum_{i,k=1}^n a_{ik} \Delta X_i \Delta X_k = 1$$

If by decreasing the gain  $W(0)$  the roots of the characteristic equation are made so small, that  $e^{\tau\lambda} \approx 1$ ,  $W(\lambda) \approx W(0)$ , then, in accordance with (15)

$$\lambda = -\frac{W(0)}{C_v^2} < 0$$

and similar slow processes of extremal control always possess monotonous stability. However, with small gains the extremal control time or the self-adjusting time is large and the errors considerable.

Errors produced by drift of the extremum with constant speed equal

$$\Delta X_k^* = -\frac{1}{\Delta(0)} \sum_{v=1}^n (-1)^{k+v} \Delta_{kv}(0) \dot{X}_{vl} \quad (16)$$

It is noted that the value

$$(-1)^{k+v} \frac{\Delta_{kv}(0)}{\Delta(0)}$$

equals the area bounded by the curved weighting function  $K_{kv}(t)$  of a closed-loop system, corresponding to the transfer function

$$\frac{\Delta_{kv}(D)}{\Delta(D)} \quad (-1)^{k+v} \frac{\Delta_{kv}(0)}{\Delta(0)} = \int_0^\infty K_{kv}(t) dt = T_{kv} \quad (17)$$

Values  $T_{kv}$  have time dimensions and will be called 'areas of weighting functions'. If the system of extremal control is in general disconnected at  $\Delta X_k^*(0) = 0$ ,  $T_0 W_{ki}^c = 0$ ,  $\delta f_{i0} = 0$ , and

$$\Delta X_k^* = -\int_0^t \dot{X}_{kl} dt = -\dot{X}_{kl} t$$

where we consider  $\dot{X}_{kl} = \text{const}$ .

With a disconnected system of extremum control, deviation  $\Delta X_k^*$  increases and in time  $T_k$  exceeds the permissible value  $\Delta X_{kg}^*$  where

$$\Delta X_{kg}^* = -T_k \dot{X}_{kl} \quad (18)$$

Time intervals  $T_k$  are called adjustment-loss times, as distinct from the general adjustment-loss time, mentioned above.

From (16) and (17) it follows that

$$\Delta X_k^* = -\sum_{v=1}^n T_{kv} \dot{X}_{vl}$$

Errors  $\Delta X_k^*$ , produced by constant drift of the extremum point in a closed system, naturally, must not exceed  $\Delta X_{kg}^*$ .

It follows from this that the weighting function areas must satisfy correlations

$$\left| \sum_{v=1}^n T_{kv} \dot{X}_{vl} \right| < T_k |\dot{X}_{kl}| = |\Delta X_{kg}^*| \quad (19)$$

Assuming in particular

$$\dot{X}_{1l} = \dots = \dot{X}_{k-1l} = \dot{X}_{k+1l} = \dots = \dot{X}_{nl} = 0$$

(moreover  $\Delta X_{vg}^*$  remain final values) one obtains

$$|T_{kk}| < T_k$$

i.e. the weighting function areas must be smaller than the adjustment-loss times.

The curtailment of the weighting function areas (decrease of static errors) may be attained by increasing the amplification. However, the increase of gains, starting with certain values, leads to the loss of stability of the extremal loop.

The increase in critical values of the gains and curtailment of times of transient processes of the extremal loop requires the diminution of transient lags  $\tau_v$  between points of action of controlling elements and measuring points of output parameters in the industrial process itself.

The control of output parameters may be realized at the output of the whole industrial process [Figure 4 (a)], at intermediate points [Figure 4 (b)], and both at intermediate points and at the output [Figure 4 (c)].

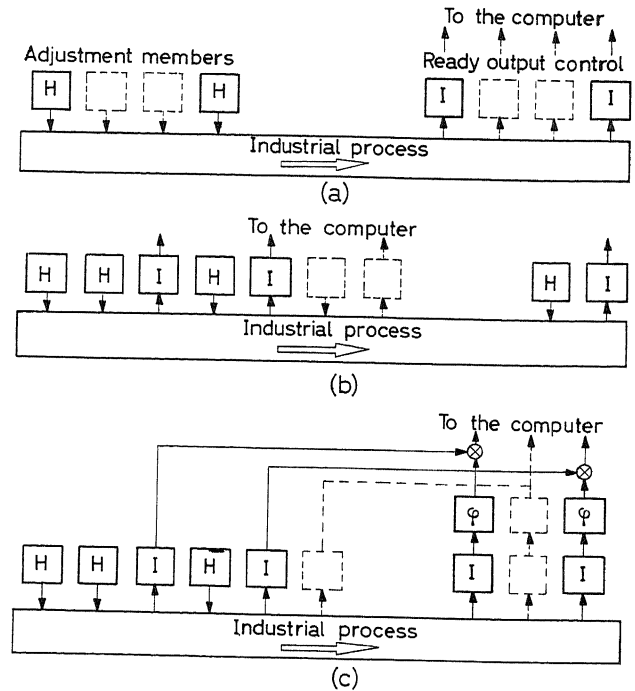


Figure 4

From the viewpoint of lag decreasing and possibility of time curtailment of transient processes, a circuit having parameter control at intermediate points has a decisive advantage over a scheme with control of final output [Figure 4 (a)] since it corresponds to the arrangement of information transmitters in the immediate vicinity of the controlling elements. However, this circuit also has one essential drawback.

The quality index extremum, calculated on the basis of measurements at intermediate points, may not correspond to the extremum of output quality at the industrial process output.

A more perfect circuit is the combined type [Figure 4 (c)] where the control at intermediate points is combined with the output parameter control at the industrial process exit. In this circuit the signals of the parameter transmitters of the finished production pass through low frequency narrow-band filters  $Q$ , for instance integrating elements, after which they are added to the signals of corresponding transmitters, controlling the output parameters at intermediate points. This circuit conserves the quick action of circuit 4 (b) and at the same time possesses the accuracy of control of slow changing parameters, near to the accuracy of control of circuit 4 (a).

However, the above extremal control system even with an improved informational section has a limited general application.

In fact, as seen from eqn (12) the dynamics of quasi-stationary processes of extremal control in this system depends on co-

efficients  $a_{iv}$  of quadratic shape of the quality index, depends on time lag  $\tau_v$ , errors of simulation of this delay  $\delta\tau_v$  and intensity of search elements.

For some industrial processes these parameters may be considered permanent, for others they are subject to comparatively slow random variations.

For the first processes the extremal control loop once adjusted maintains its efficiency for a long time. In the second case the extremal control loop itself needs periodic adjustment, accomplished by changing the transfer functions  $W_{kv}(D)$  of the filters or just their gains  $W_{kv}(0)$ , and also, perhaps, the time delays.

The necessity of adjustment arises due to the fact that, even though the stability of slow processes of the extremal control is maintained in a wide range of variations  $a_{i\nu}$ ,  $R_{\nu}(\delta\tau_{\nu})$ , a guarantee of the necessary quality of the extremal control is possible only by a suitable selection of transfer function filters.

To this must be added, that even in those processes where parameters  $a_{iv}$ ,  $\tau_v$ ,  $R_v(\delta\tau_v)$  remain unchanged the initial adjustment requires either *a priori* knowledge of these parameters, or their experimental determination.

An increase of the chance of general acceptance of extremal control systems with continuous industrial processes, in the sense of decrease of the necessary *a priori* information, may be attained at the expense of parametric extremum adjustment of the basic extremal loop.

### Self-Adjusting System with Parametric Extremum Adjustment of the Basic Loop

To realize an extremum adjustment of the basic control, it is desirable to select the adjustment parameters of this loop and the quality index so that the latter shall be the only extremum in the working portion of possible variations of adjustment parameters.

As an adjustment index of the basic loop, it is natural to select the mean value or, more accurately, the mathematical expectation of the same production quality index  $Q$ , which is utilized in the basic extremal loop.

Moreover, in accordance with (2)

$$M[Q] = Q_t + \frac{1}{2} \sum_{i,j=1}^n a_{ij} M[\Delta X_i(t - \tau_i) \Delta X_j(t - \tau_j)] \\ + M[\delta Q_f]$$

it is possible to show, if the errors of simulation lag are so restricted, that

$$R_v(\delta\tau_v) > 0 \quad (v=1, 2, \dots, n)$$

$$\begin{vmatrix} d_{11} & \dots & d_{1n} \\ \dots & \dots & \dots \\ \dots & \dots & \dots \\ d_{nt} & \dots & d_{nn} \end{vmatrix} > 0 \quad (20)$$

where  $d_{y,k} = W_{y,k}(0)$ .

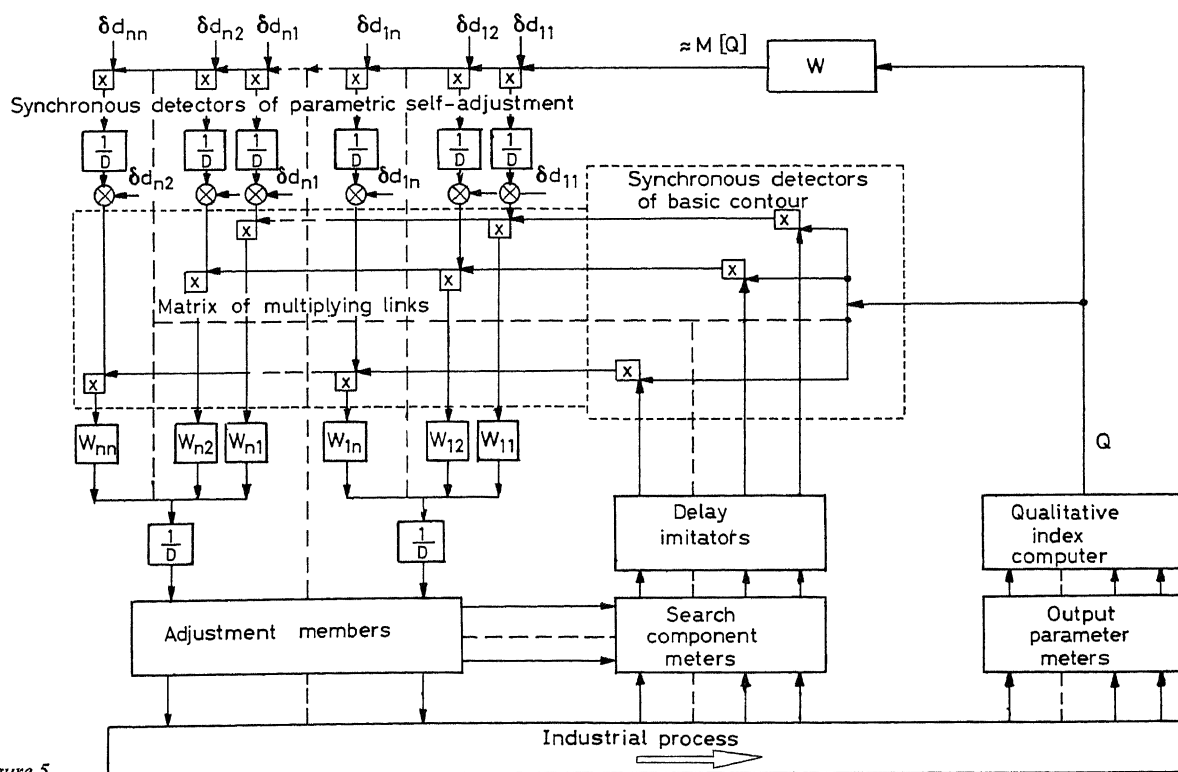
Then the estimation  $M[Q]$  always has an extremum by the adjustment parameters  $d_{vk}$  whereupon this extremum is the only one in region (20).

Taking into account the availability of the single extremum by the adjustment parameters and the general principle of extremal control, it is easy to lay out a control system of an industrial process with self-adjustment of the basic loop.

This diagram is shown in *Figure 5*. The basic loop of the extremal control of the industrial process is here similar to the one previously examined. The difference is only in the presence of multiplier 'matrix' links, which realize varying transfer numbers  $d_{ik}$ .

Besides the basic loop the system has a loop of extremum adjustment of the transfer numbers matrix.

It is assumed that the transfer numbers  $a_{i,v}$  of the indus-



*Figure 5*

trial process change slowly in time even as compared with quasi-stationary processes of the basic extremal loop. More accurately, it is assumed, that the time of substantial change of the transfer numbers  $a_{iv}$  is considerably greater than the general adjustment-loss time  $T$  (see above). It is noted that, in principle, forced self-adjusting processes of transfer numbers are also possible. However, the dynamics of forced processes is complex, and for their organization it is not sufficient to only have the existence of an extremum of appraisal  $M[a]$  by the transfer numbers of the basic loop. Thus, as a typical regime of the system operation with two extremal loops, a regime with the following grading of process flow speeds is assumed: (a) Search elements in the basic loop (the frequency processes); (b) working processes in the basic loop; (c) search of oscillations in the loop of extremal self-adjustment of the transmitting numbers, and (d) working processes of the extremal self-adjustment of transfer numbers (the frequency processes).

With the above grading of process flow speeds, both the processes in the basic loop, as well as the processes in the self-adjusting loop of transfer numbers, are quasi-stationary. The dynamics of working processes, moreover, are near to the dynamics of ideal gradient systems (4). From this position and presence of an extremum of transfer numbers  $d_{vk}$  it follows, that upon fulfilment of weak conditions (20) a quasi-stationary process of self-adjustment of the basic loop is always stable.

The above control system has considerable universal acceptance. By joining it with a plant (industrial process) with few known characteristics, the system matches automatically transfer numbers corresponding to the quality index extremum in the framework of the given structure of the system.

A further increase in 'flexibility' or universality of the system is made by introducing extremal adjustment of the delay simulator, extremal adjustment of the filter time constants and others. However, all of this involves further complexity of the system.

### The Possibility of Extremal Control of Non-automated

The main difficulties in introducing extremal control of industrial processes at the present stage are connected with the

complete automation of output parameter control and complete automation of machine adjustment, securing the industrial process. The technology of most industrial processes even continuous, did not yet reach the level at which it is possible to achieve continuous automatic control and adjustment. Therefore, it is of great interest to find the means of extremal control for discrete semi-automatic or hand control and adjustment.

The general algorithm of the extremal control and the computing section of the control system may, moreover, remain the same, as in a continuous automatic system.

Estimation of the information output capacity of the measuring points, necessary for transmission of the search elements, indicates that for processes with considerable adjustment-loss times a non-automated control is possible. For such processes, a periodic hand adjustment of machines is also possible, which guarantees the industrial process taking place. In this case the output signals of synchronous detectors are transmitted to the integrated indicators. Operators guided by the indicators of these devices, periodically correct the adjustment of the machines. With this type of organization of the extremal control the control system itself becomes a computer, either digital or analogue, equipped with input and output arrangements. The closing of the loop of an extremal control is here accomplished by human supervisors and operators.

### References

- <sup>1</sup> KRASOVSKI, A. A. *Some Conditions of Application of Self-adjusting Control Systems with Continuous Industrial Processes*. No. 1. 1961. Moscow; Izvestia Academy of Sciences U.S.S.R.
- <sup>2</sup> COWDEN, D. T. *Statistical Methods in Quality Control*. 1957.
- <sup>3</sup> PERLMAN, I. I. Statistical automats, relay type and some methods of their investigation. *Automat. Telemekh.*, Moscow 22 (1961) 6
- <sup>4</sup> KRASOVSKI, A. A. *Dynamics of Continuous Systems of Extreme Regulation, Based on Gradient Method*. No. 3. 1959. Moscow; Izvestia Academy of Sciences U.S.S.R.

### DISCUSSION

P. M. E. M. VAN DER GRINTEN, *Central Laboratory, Staatsmijnen, Geleen, Netherlands*

For various reasons you recommend the use of high-frequency search oscillations. This implies that the phase-lag in the delay line applied will increase and a small error in the time delay will cause large errors in the computations. Apart from this there will probably be dynamic effects in the process, other than pure delay, which also affect the amplitudes for high frequencies. Might these complications have an adverse effect on the practical applications of the proposed method?

A. A. KRASOVSKI, *in reply*

High-frequency oscillations are understood here to be those having a high frequency as compared with the self-adaptive processes. As compared with a delay, the periods of these oscillations can be large. For cases where this does not take place and the error of delay simulation is larger than the period (correlation time) of search oscillations, it is proposed to introduce self-adjustment for time delay of the model (simulator). The dynamic effects are taken into account in the system analysis.

# Principle and Application of an Extremal Computer

R. PERRET and R. ROUXEL

## Summary

This paper deals with the structure of a special purpose computer for the automatic search of the optimum operating condition of a process consisting of one time constant and a delay. The structure proposed is independent of the dynamics of the system and it is not necessary to identify the process dynamics in order to apply this method.

The first part of the paper is devoted to the theoretical considerations leading to the establishment of the proposed control law and its subsequent experimental verification utilizing an analogue computer. The second part of the paper describes the application of this method for the minimization of the ohmic loss of an alternator coupled to a power network. This experimental investigation permits confirmation of the results obtained from a theoretical study and helps to improve the design of the computer.

In particular, this investigation demonstrates: (a) the possibility of building a simple computer for the search of an extremum using analogue and sequential circuits; (b) the self-adaptive properties of the computer even in the presence of fast perturbations and the stability of the method; (c) the possibility of compensating pure time delay (dead time) present in physical systems; and (d) the importance of filtering in the practical application of this method.

## Sommaire

Cette étude propose une structure de calculateur destiné à la recherche automatique d'un optimum de fonctionnement pour un processus doué d'une constante de temps et d'un retard. On doit noter que cette structure est indépendante de la nature de la dynamique du système commande et que l'application de cette méthode n'implique pas la détermination de cette dynamique.

Une première partie est consacrée à la mise au point et à l'expérimentation sur un calculateur analogique de la loi de commande proposée. Une deuxième partie décrit l'application de cette méthode à la minimalisation des pertes ohmiques dans un alternateur couplé au réseau. Cette étude expérimentale a permis, d'une part de confirmer l'étude théorique, d'autre part d'apporter des perfectionnements pratiques au calculateur étudié.

Cette étude montre en particulier: (a) qu'il est possible de réaliser selon des structures simples, analogique et séquentielle, un calculateur de recherche d'extremum; (b) le caractère d'auto-adaptation, en particulier vis-à-vis de perturbations rapides, et de haute stabilité de la méthode; (c) la possibilité de compensation des retards propres à tout système physique; et (d) l'importance du problème de filtrage pour la mise en application du procédé.

## Zusammenfassung

In der vorliegenden Untersuchung wird der Aufbau eines speziellen Rechengerätes diskutiert, das automatisch das Optimum eines Prozesses sucht, der sowohl eine Verzögerung 1. Ordnung als auch eine reine Totzeit besitzt. Hervorzuheben ist, daß der Aufbau des Rechners unabhängig von der Art des dynamischen Verhaltens des geregelten Prozesses ist; zur Anwendung der vorgeschlagenen Methode ist es nicht notwendig das dynamische Verhalten des Prozesses zu ermitteln.

Der erste Teil der Untersuchung schildert die Herleitung des vorgeschlagenen Optimierungsgesetzes sowie experimentelle Untersuchungen mit Hilfe eines Analogrechners. Im zweiten Teil wird die Anwendung der Methode zur Minimierung der ohmschen Verluste in einem mit dem Netz gekoppelten Generator beschrieben. Die hier

gewonnenen experimentellen Ergebnisse haben einerseits die theoretischen Untersuchungen bestätigt, andererseits praktische Hinweise für den Bau des Rechengerätes geliefert.

Die vorliegende Untersuchung zeigt unter anderem: a) Es ist mit einfachen analogen und sequentiellen Schaltungen möglich, ein Rechengerät zu bauen, das selbsttätig die Extremwerte sucht. b) Die Eigenschaft der Selbsteinstellung des Gerätes, auch bei Vorhandensein schnell veränderlicher Störgrößen, sowie die gute Stabilität des Verfahrens. c) Die Möglichkeit, die in physikalischen Systemen vorhandene Verzögerung (Totzeit) auszugleichen. d) Die Bedeutung des Filterproblems bei der praktischen Anwendung der Methode.

## Introduction

Many industrial processes have an optimum working region which can be expressed as an extremal value of a figure of merit  $f(x_i, p_k)$ . This figure of merit depends upon the acting variables  $x_i$  on the one hand and the disturbing values  $p_k$  on the other; the latter coming either from variations in the system environment or from the system itself. These disturbances act slowly with respect to the dynamics of the system and are likely to modify the state of the system very appreciably. In the present study a method is established of controlling the acting variables from a measure of the figure of merit. The function of this closed-loop control is to compensate for the action of the disturbing values which tend to shift the system from its optimum working condition. Moreover, this servo should be able to seek the optimum from any initial conditions. In order to obtain this double objective, measurements are made in the process which allow the figure of merit to be determined more or less directly, and acting elements to modify the process are installed. In general the disturbing values are inaccessible, and all that is known about the function  $f$  is that it has an extremum. This function is usually only available after a dead time or time constant which are inherent in all physical systems. This means in fact that only the function  $h(x_i, p_k)$  is available, connected to  $f(x_i, p_k)$  by a differential relationship.

The following remarks can be made about this type of closed-loop system. The study of a system whose performance is determined by several independent acting variables can be considered as a system with one variable if at any given instant only one variable is controlled. The control is switched successively from one variable to another. Under these conditions it is necessary only to study a system with one acting variable, which is designated  $x$ . It is necessary to modify the acting variable  $x$  with a speed  $dx/dt$  whose sign and amplitude are determined.  $x$  itself cannot, in fact, be controlled as in a normal servo system since the reference value which corresponds to the extremum is unknown. This remark makes apparent the adaptive character of the control envisaged, since the role of this control is to make the system tend towards an optimum which is unknown and possibly variable.

This study has led to consideration of certain special cases which are likely to complete the work undertaken in the same spirit by other workers<sup>1-8</sup>. The study has been done theoretically by analogue methods and then experimentally on an alternator connected to a supply.

### Analogue Study

First, the system to be studied is defined from a static and dynamic point of view; the method of control is then envisaged, particularly the nature of the necessary switching. The study of limit cycles enables definition of ideas of stability of the optimum. Finally, the influence of the various control parameters is studied.

### Description of the System

The system studied is represented schematically in a simplified form in *Figure 1*. The controlled system comprises the elements *B* and *T*, where *B* represents the function to be optimized, here supposed to be a minimum. Near the minimum the function  $f(x)$  can be considered as  $\alpha x^2$  ( $\alpha > 0$ ) taking the origin as the abscissa of the minimum. The element *T* represents a time constant  $\tau$  which can be due either to the system itself or to the measuring device. The integrator *A* does not, strictly, belong to the system but to the controller; but it can be considered as inherent in all control devices, and it is convenient to connect it artificially to the system. The element *R* represents a pure dead time  $\tau_r$ . It is possible to lump at the input the sum of the lags which appear at any point in the chain. These lags intervene then simply as a retarded action.

The effect of a time constant  $\tau_e$  at the input can easily be compensated since there is the derivative of the acting variable. It is only necessary to replace the integrator with an element of transfer function  $(\tau_e + 1/p)$ . Hence the general system to be considered is that shown in *Figure 2*, which is easily derived from *Figure 1*. Finally, a constant amplitude *K* is imposed on the derivative of the acting variable.

The object of the control is to determine the sign of  $dx/dt$  knowing  $h[x(t)]$ . The relationships defining the dynamic functioning of the system can be written

$$\tau \frac{dh}{dt} + h = \alpha x^2 \quad (1)$$

$$\frac{dx}{dt} = \varepsilon K \quad (\varepsilon = \pm 1) \quad (2)$$

The initial conditions fix the values  $x(0)$  and  $h[x(0)]$ . Considering  $x$  as the independent variable and introducing the dimensionless variables defined by

$$X = \frac{x}{K\tau} \quad (3)$$

$$H = \frac{h}{2\alpha K^2 \tau^2} \quad (4)$$

the solutions for the system are

$$H = \frac{X^2}{2} - \varepsilon X + 1 + \lambda e^{-\varepsilon X} \quad (5)$$

from which

$$\frac{dH}{dX} = X - \varepsilon - \varepsilon \lambda e^{-\varepsilon X} \quad (6)$$

$$\frac{d^2 H}{dX^2} = 1 + \lambda e^{-\varepsilon X} \quad (7)$$

and

$$H = \frac{X^2}{2} - \varepsilon \frac{dH}{dX} \quad (8)$$

$$H = \frac{X^2}{2} - \varepsilon X + \frac{d^2 H}{dX^2} \quad (9)$$

$\lambda$  is a dimensionless coefficient defined by the initial conditions.

Consider the evolution of the system in the  $H(x)$  plane. This representation is useful since eqn (2) implies that  $X$  is a linear function of time. The dynamics of the system are thus represented in this plane by two networks of curves, symmetrical about the vertical axis, which correspond to the two possible values of  $\varepsilon$ . These curves are represented in *Figure 3* for the case  $\varepsilon = -1$ . Also drawn in *Figure 3* are

$$(a) \quad \frac{dH}{dX} = 0, \text{ or } H = \frac{X^2}{2} \quad \text{from eqn (8)}$$

this is a parabola representing the static state *F*

$$(b) \quad \frac{d^2 H}{dX^2} = 0, \text{ or } H = \frac{X^2}{2} - \varepsilon X \quad \text{from eqn (9)}$$

These are two parabola  $H$  symmetrical with respect to the vertical axis; the parabola relative to  $\varepsilon = -1$  is designated  $H^-$ .

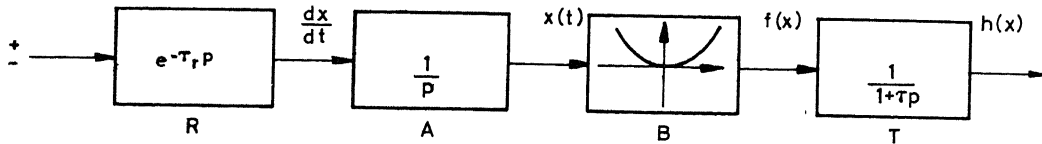


Figure 1. Block diagram of the system

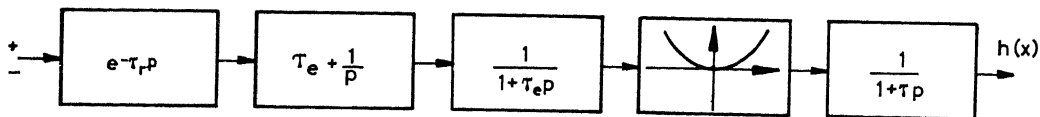
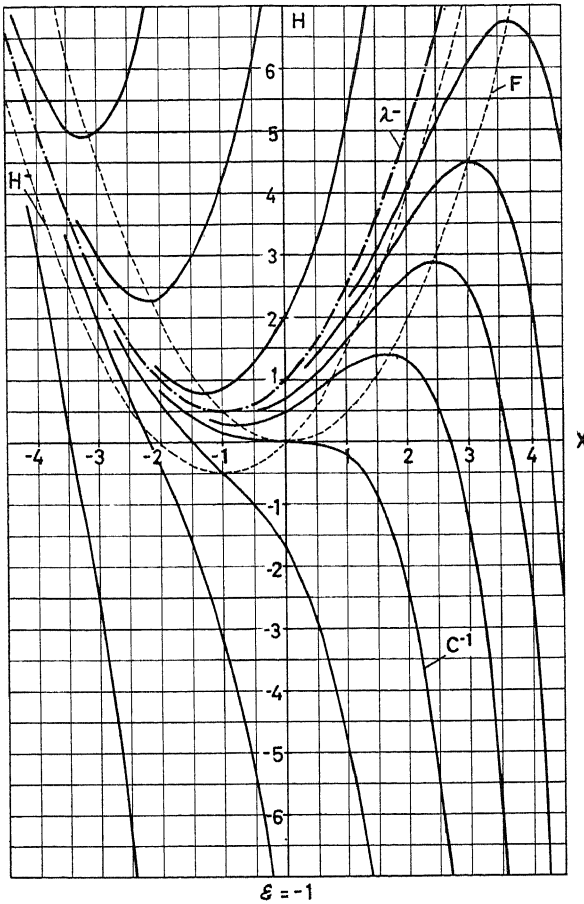


Figure 2. Equivalent diagram




 Figure 3. Curves  $H(x)$  for  $x$  decreasing

These curves also correspond to the zero values of  $dh/dt$  and  $d^2h/dt^2$ .

The trajectories are asymptotic to the symmetrical parabolas of equation

$$H = \frac{X^2}{2} - \epsilon X + 1 \quad (10)$$

Switching the acting variable causes the point representing the dynamic state of the system to pass from one network of curves to the other.

#### Switching

**Proposed Law**—The proposed method of searching for the minimum consists in acting at constant speed; positive or negative, on the acting variable  $x$  according to eqn (2); that is, to impose a positive or negative sign on the derivative  $dx/dt$ . The choice of sign is determined by a switching device which operates according to the following sequence.

The sign of  $\epsilon$  is modified each time that the quantity

$$\frac{d^2h}{dt^2} - \delta = g$$

or

$$G = \frac{d^2H}{dT^2} - \Delta$$

with

$$\Delta = \frac{\delta}{2\alpha K^2}$$

passes from a negative to a positive value (switching  $N$  or 'natural'), and this switching is to continue with a period  $\tau_w$  ( $W = \tau_w/\tau$ ) as long as the function  $G$  remains positive (switching  $F$  or 'forced').

$\delta$  is a positive or negative threshold. A zero value of  $\delta$  with the proposed switching law leads to a search for the minimum value of  $dh/dt$ ; in other words, to make  $h$  tend to its minimum value as quickly as possible.

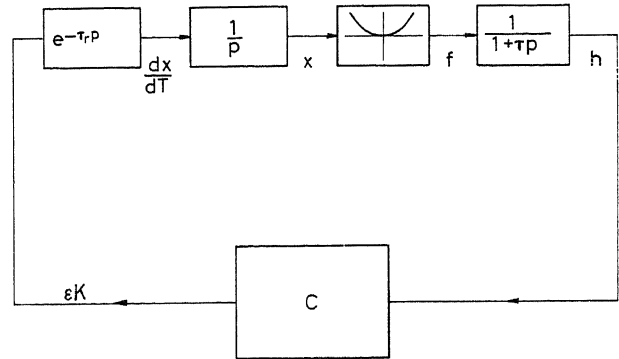


Figure 4. Block diagram of the closed-loop system

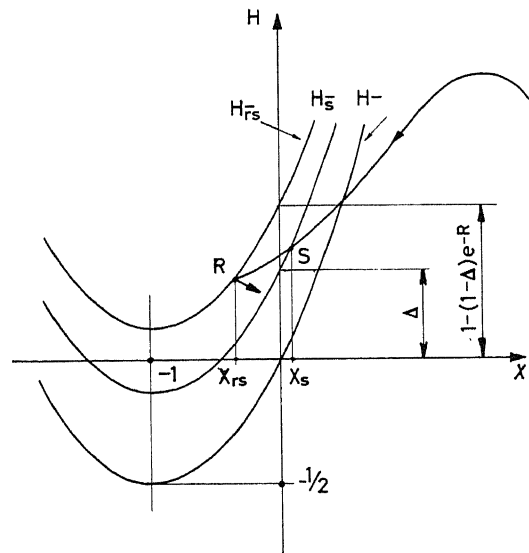


Figure 5. Switching processes with a threshold and lag

The forced switching has been introduced both to take account of part of the unfavourable initial conditions which do not allow the system to switch naturally, and to avoid instability in the case where, upon switching, the system diverges ( $G > 0$ ).

The switching process can now be defined for a system with dead time (Figure 4).  $C$  represents the switching device which produces the law defined above.

Consider (Figure 5) the case of one curve ( $\epsilon = -1$ ). The quantity

$$G = \frac{d^2H}{dT^2} - \Delta$$

which should start the natural switching action becomes zero at  $S$  when

$$1 + \lambda e^{-\epsilon X_s} = \Delta \quad \text{from eqn (7)} \quad (11)$$

Switching is only effective on  $dx/dt$  at a time  $\tau_r$  later. There results the relationship between the abscissa  $X_{rs}$  of actual switching and  $X_s$  of starting

$$X_{rs} - X_s = \varepsilon R \quad (12)$$

$R$  is the dimensionless lag  $\tau_r/\tau$ .

Lastly the point  $P$ , being on the considered trajectory, must satisfy eqn (5)

$$H_{rs} = \frac{X_{rs}^2}{2} - \varepsilon X_{rs} + 1 + \lambda e^{-\varepsilon X_{rs}} \quad (13)$$

The elimination of  $\lambda$  and  $X_s$  between eqns (11) to (13) leads to the equation for the actual switching points

$$H_{rs} = \frac{X_{rs}^2}{2} - \varepsilon X_{rs} + 1 - (1 - \Delta) e^{-R} \quad (14)$$

### Properties

1. The switching points  $H_{rs}$  described by eqn (14) are parabolas. When  $\Delta$  or  $R$  vary, these parabolas keep the same axis of symmetry  $X = \varepsilon$ . They are derived from parabolas  $H$  by a vertical shift of  $1 - (1 - \Delta) e^{-R}$ . If there is no dead time, natural switching occurs on parabolas  $H_s$  whose equation is

$$H_s = \frac{X_s^2}{2} - \varepsilon X_s + \Delta \quad (15)$$

If there is no threshold or dead time, natural switching occurs on parabola  $H$  given by

$$H = \frac{X_s^2}{2} - \varepsilon X_s \quad (16)$$

2. When  $R$  is different from zero, the switching on  $H_{rs}$  occurs only if switching action has started on  $H_s$  beforehand. If not curve  $H_{rs}$  can be crossed without natural switching [Figures 12 (c) and 13 (c)].

3. A threshold  $\Delta$  equal to unity is equivalent to an infinite dead time. The switching parabolas  $H_{rs}$  and  $H_s$  tend to the asymptote parabola  $\lambda$ , switching cannot occur, and the system diverges. Hence the condition  $\Delta < 1$  must be observed.

4. The switching mechanism described is useful only if the period of forced switching  $W$  is greater than the lag  $R$ . In fact the control principle itself implies the observation of the sign of  $G$ . In the presence of a dead time, the start of natural switching action causes, in principle, the start of a forced switching action which will occur in all cases where  $W$  is less than  $R$ . Hence the condition  $W > R$  must be observed to avoid false switching.

5. The two parabolas  $H_s$  divide the plane into four regions (Figure 6). In zone I,  $G$  is negative; no switching can occur in this zone. The dynamic curves of the two networks are increasing, and all cross a parabola  $H_{rs}$  on which a switching takes place [Figure 15 (e)].

In zone II,  $G$  is positive; hence the production of periodic forced switching as long as the system remains inside this zone [Figure 15 (d)].

All curves of the two networks are decreasing and finally enter the zones III and IV [Figure 15 (c)]. The final state should establish itself, at least partially, in the zones III and IV where

natural or forced switching leading to limit cycles can occur, depending on the initial conditions.

6. Eqns (14) and (16) suggest the possibility of compensating for a dead time by a negative switching threshold. If the equation for compensation is produced

$$1 - (1 - \Delta) e^{-R} = 0$$

or

$$\mu = 1 \quad (17)$$

where

$$\mu = (1 - \Delta) e^{-R}$$

the switching curve  $H_{rs}$  becomes the curve  $H$  of switching in the absence of threshold  $\Delta$  and dead time  $R$  (Figure 5). This remarkable result is verified on the analogue computer [Figures 12 (e) and 13 (e)]. The sum of these properties justifies the proposed switching law.

### Limit Cycles

The application of constant, non-zero, positive or negative acting variable leads, in the absence of disturbances, to stable periodic oscillations. Two principal types of limit cycle can be observed in the plane  $H(X)$ :

type NN where only natural switching occurs,

type NF where natural and forced switching alternate.

Figure 7 shows these two types of cycle in the presence of dead time and a threshold.

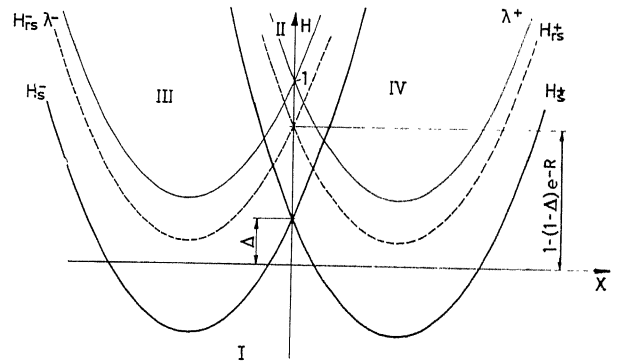


Figure 6. Switching zones

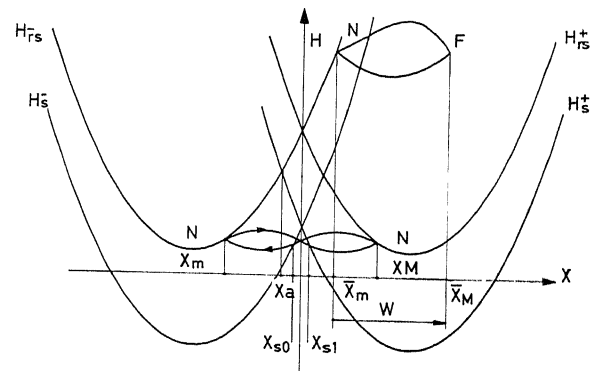


Figure 7. Limit cycles

*Calculation of Limit Cycles NF*—Calling  $\lambda_0$  and  $\lambda_1$  the two values of  $\lambda$  relative to the two trajectories which constitute the limit cycle, the following equations are obtained.

$$\begin{aligned} 2X_m &= \lambda_1 e^{-X_m} - \lambda_0 e^{X_m} & [\text{from eqn (13)}] \\ 2X_M &= \lambda_1 e^{-X_M} - \lambda_0 e^{X_M} & [\text{from eqn (13)}] \\ 1 + \lambda_0 e^{X_s} &= \Delta & [\text{from eqn (11)}] \\ X_s - X_m &= R \\ X_M - X_m &= W \end{aligned} \quad (18)$$

$X_s$  corresponds to the crossing of  $H_s$ .

Values can then be deduced for

$$X_m = \frac{\mu(e^{2W} - 1) - 2We^W}{2(e^W - 1)} \quad (19)$$

$$X_M = \frac{\mu(e^{2W} - 1) - 2W}{2(e^W - 1)} \quad (20)$$

Figure 8 shows the variation of  $X_m$  as a function of  $W$  and  $\mu$ .

*Calculation of Limit Cycles NN*—Similar equations to eqn (18) show that in this case

$$X_M = -X_m \quad (21)$$

and

$$X_M = -\frac{\mu}{2}(1 - e^{2X_M}) \quad (22)$$

The cycle is symmetrical, and its amplitude is given by solution of eqn (22). The amplitude is independent of  $W$ , Figure 9 showing how it depends upon  $\mu$ .

*Limiting Case*—The curves of Figures 8 and 9 can be used only if the type of limit cycle considered is known beforehand. The

limiting case which separates cycles NN and NF is that where the point  $X_m$  lies on  $H_s$ ; in other words, when the following equation must be added to those of eqns (18)

$$1 + \lambda_1 e^{-X_m} = \Delta \quad (23)$$

From these can be deduced eqn (24) which exists between the parameters  $R$ ,  $\Delta$  and  $W$  in order that the limit of the appearance of a cycle NN may be reached

$$R = \log \left[ \frac{(1 - e^W)(1 - \Delta)}{(1 - e^{-W})(1 - \Delta) - 2W} \right] \quad (24)$$

This expression leads to the plotting of the network of curves of Figure 10.

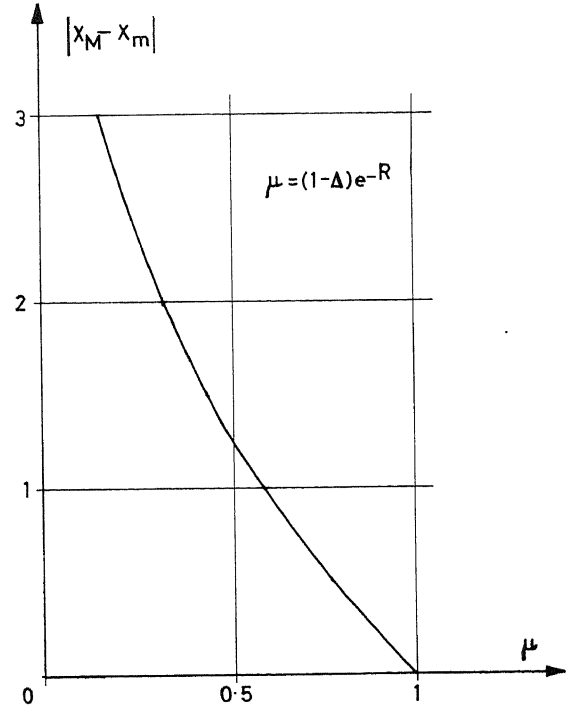


Figure 9. Limit cycle (NN)

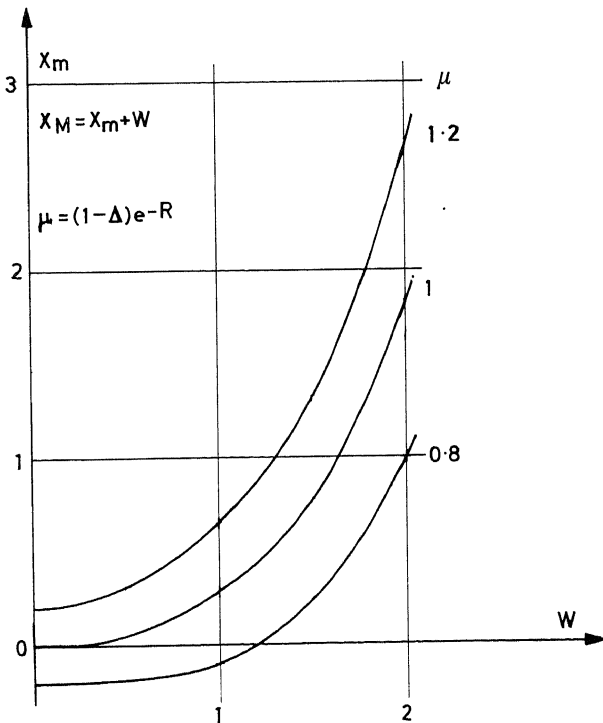


Figure 8. Limit cycle (NF)

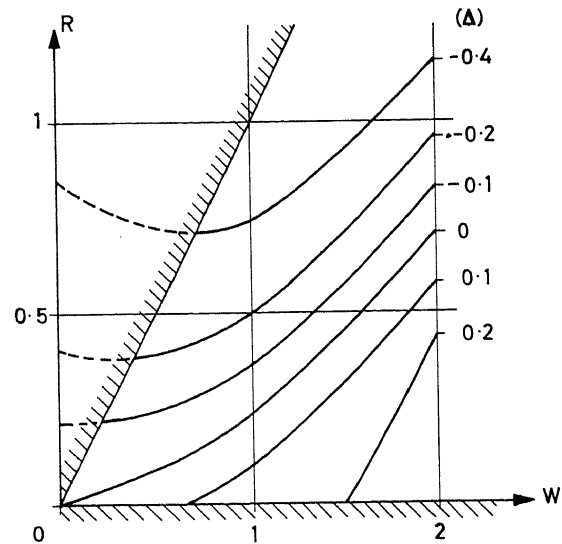


Figure 10. Limiting cases of the occurrence of limit cycles NN and NF

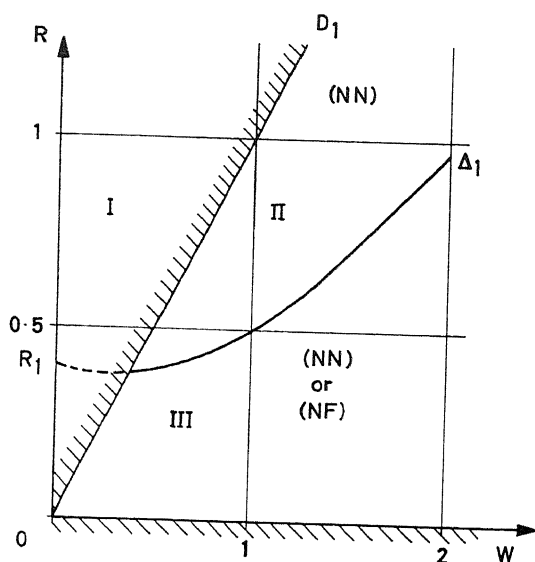


Figure 11. Limiting cases of the occurrence of limit cycles NN and NF

Three zones are indicated in Figure 11. Proper operation is not possible in zone I since  $R > W$ . Zone III leads to limit cycles NN or NF, depending on the initial conditions, and in zone II all limit cycles are NN. In zone II the interest of oscillations NN is that use can be made of their symmetry to determine the optimum state of the system and to fix it in this position. Such a realization is described in the experimental section.

#### Influence of the Parameters

**The Nature of the Parameters**—Two sorts of parameter can be distinguished: first, those connected with the system and which cannot be acted on. These are the time constant  $\tau$  and the dead time  $\tau_d$ . The time constant  $\tau$  occurs as a normalized value by means of the product  $K\tau$ , while the dead time is defined by the dimensionless variable  $R$  proportional to the real dead time of the system.

The second type of parameter is connected with the controller and can be acted on. The dimensionless switching time  $W$  which is proportional to the real period  $\tau_w$ , the gain  $K$ , considered as a real value and the dimensionless threshold  $\Delta$  with value  $\delta/(2 \propto K^2)$ .

In practice  $\Delta$  and  $R$  intervene via the parameter

$$\mu = (1 - \Delta)e^{-R}$$

Two identical values of  $\mu$  correspond, all other things being equal, to two identical system behaviours [Figures 12 (e) and 13 (e)], if Property 2 is satisfied. Figure 13 (c), to be compared with Figure 12 (c), illustrates a case where this condition is not fulfilled.

The influence of these different parameters has been shown on the analogue computer. The recordings of Figures 12 to 15, where the numerical values are given in dimensionless form, lead to the following conclusions.

#### Results

**Oscillations NN**—These oscillations are symmetrical with respect to the optimum [Figure 12 (a)]. Their amplitude is

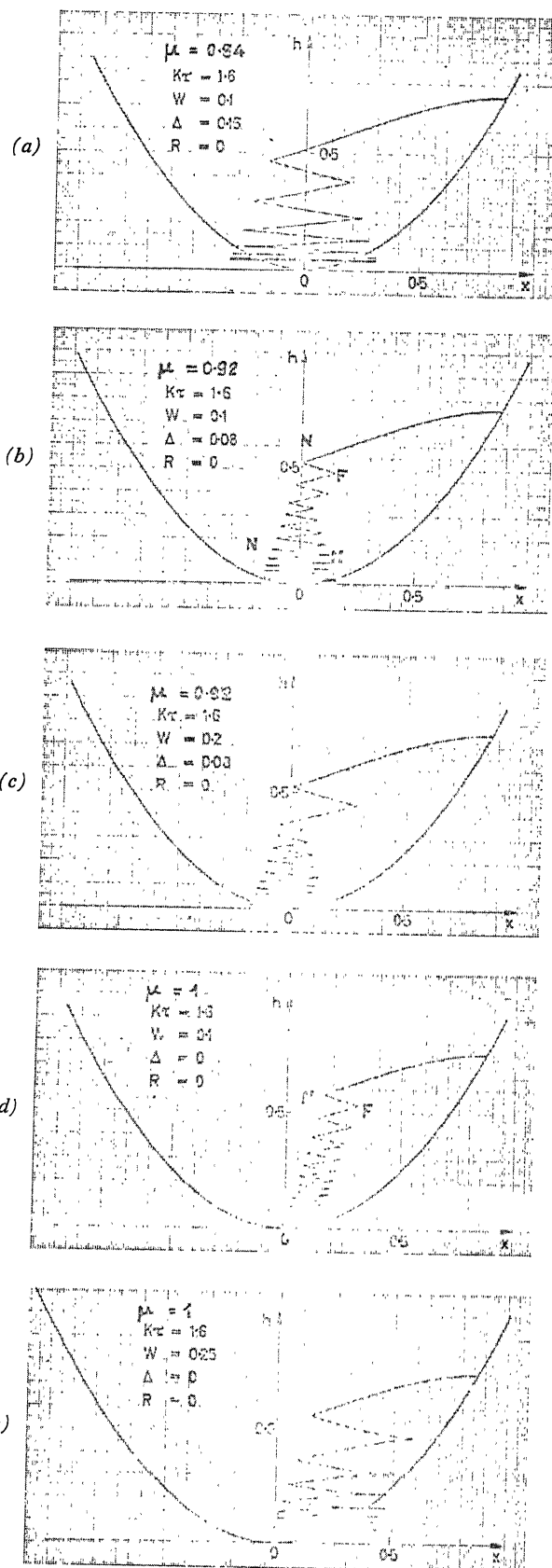


Figure 12. Influence of the parameters on the behaviour of the system

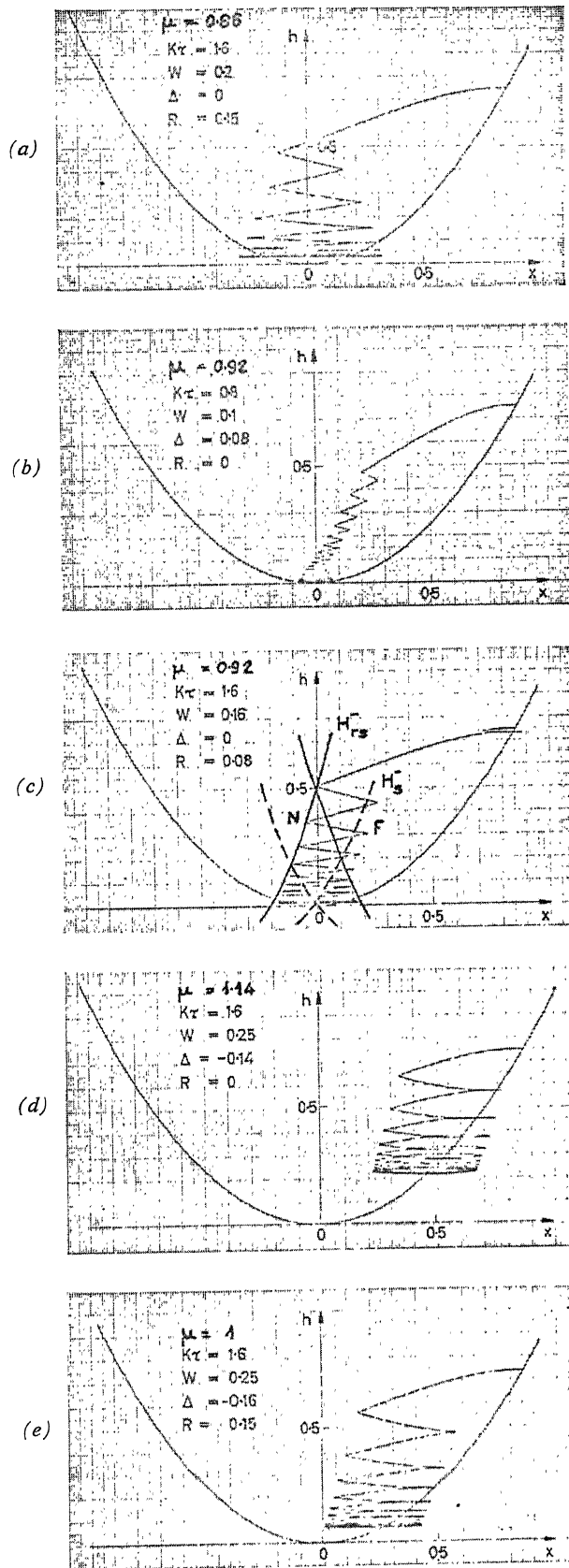


Figure 13. Influence of the parameters on the behaviour of the system

independent of  $W$  [Figure 12 (b) and (c)], proportional to  $K$  [Figures 12 (b) and 13 (b)] and reduces when  $\mu$  increases [Figure 12 (a) and (b)].

**Oscillations NF**—The average abscissa of these oscillations is further away from the optimum as  $\mu$  and  $W$  increase [Figures 13 (d) and (e); 12 (d) and (e)]. The amplitude of these oscillations is independent of  $\mu$  [Figure 13 (d) and (e)], proportional to  $W$  [Figure 12 (d) and (e)] and to the gain  $K$  [Figures 12 (b) and 13 (b)].

**Settling Time**—The settling time is proportional to the time constant  $\tau$ . An increase in gain reduces slightly the settling time [Figure 14 (a) and (b)]. In practice this time is of the order of magnitude of the time constant  $\tau$  of the system.

**Accuracy**—The accuracy is defined by the average ordinate of the limit cycle; that is, by the amplitude of cycles NN [Figures 12 (b) and 13 (b)] or the average abscissa of cycles NF [Figure 12 (d) and (e)].

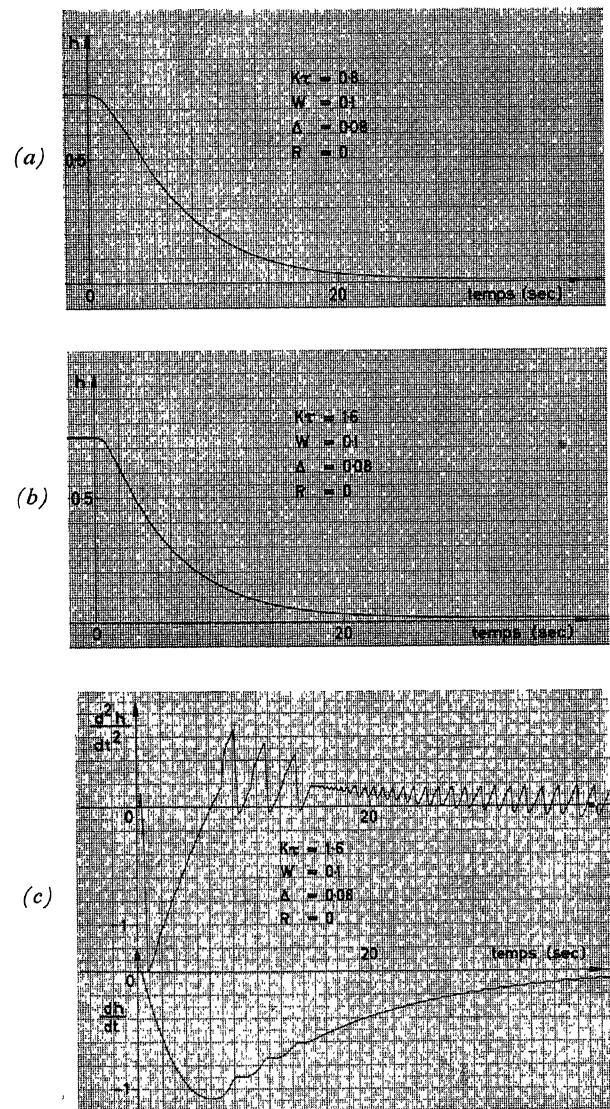


Figure 14. Influence of the parameters on the behaviour of the system

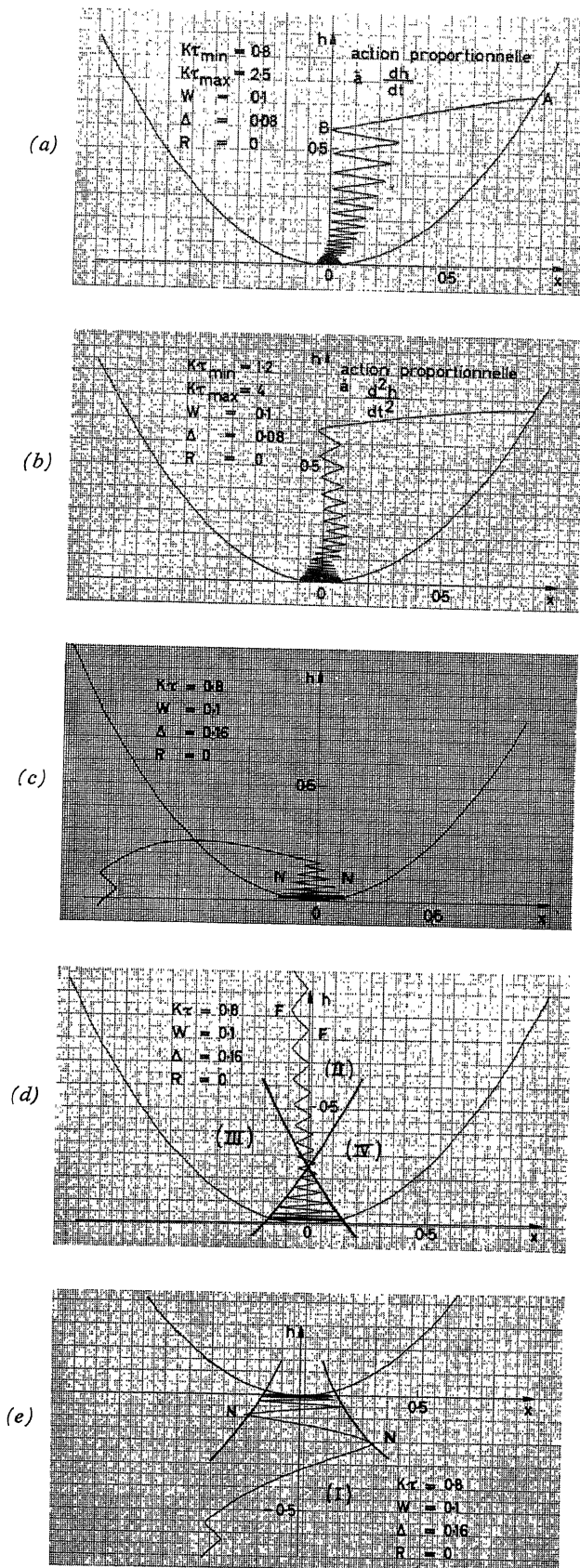


Figure 15. Influence of the parameters on the behaviour of the system

### Choice of Parameters

The forced switching period is made as small as possible in order to reduce the settling time of the system when it follows an unfavourable trajectory, and afterwards to allow the quickest possible compensation for disturbances. In the absence of a dead time, this period can be of the order of a twentieth of the system time constant  $\tau$  ( $W = 0.05$ ). With a dead time,  $W$  is chosen to be slightly greater than this dead time.

The gain must be as high as possible to reduce the settling time and to minimize the effect of disturbances. Its upper limit is set by the amplitude of the oscillations.

The threshold is chosen (Figure 16) to compensate for the lag ( $\mu = 1$ ). In practice the threshold will thus be negative, which assumes the condition  $\Delta < 1$  whatever the value of  $\alpha$ , which defines the unknown static curve.

### Experimental Study

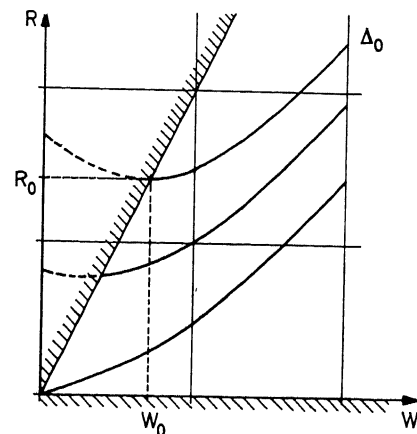
#### Statement of the Problem

The proposed method has been tested experimentally on an alternator connected to a mains supply with the object of optimizing the reactive power produced. It is known that for a given active power, the reactive power, and consequently the current produced, passes in the static state, through a minimum for a given value of excitation current. This minimum is displaced when the active power varies as is shown in Figure 18 (a).

This point corresponds to the minimum of the ohmic losses in the alternator, and can be considered as a point of optimum operation. The non-linear characteristics are not known beforehand. The main disturbances are the variations of active power. Dynamically there is a preponderant time constant (5 sec) introduced by the detection of the current supplied by the alternator. Hence the dynamics of the alternator itself do not intervene to determine the order of the system. The dead time also is negligible.

#### Brief Description of the Experimental System

The block diagram is shown in Figure 17. The alternator *I* used has a power of 6 kVA and is driven by a Ward-Leonard group which supplies the active power. The detection 2 of the current supplied by the alternator is made on the three phases by current transformers. The excitation of the alternator is

Figure 16. Choice of parameters  $W_0$  and  $\Delta_0$  for a lag  $R_0$

regulated by two silicon controlled rectifiers 14 whose gates are connected to the output  $S$  of the optimizing computer. The time constant of the excitation, of the order of 0.5 sec, can be compensated for by using the fact that the computer imposes the derivative  $dx/dt$  of the acting variable.

#### Operations Performed by the Computer—Results Obtained

The computer used for the optimal control of the alternator reactive power comprises analogue elements (operational d.c. amplifiers) to produce the algebraic and differential operations, and sequential static circuits for the logic operations.

This computer is designed first to demonstrate the principles of the method proposed in the theoretical part, and secondly to bring certain improvements, especially to suppress the oscillations around the optimum working point.

**Analogue Elements**—They realize the following functions:

(1) *First and second derivatives*. These are produced by two operational amplifiers 4 and 5 (Figure 17) which ensure at the same time a suitable filtering for the signal.

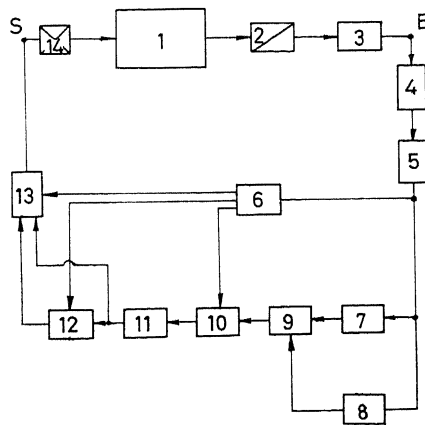


Figure 17. Extremal control of the reactive power of an alternator connected to a supply. Block diagram of the installation

1 Alternator	2 Current detector
3 Filter	4 First derivative
5 Second derivative	6 Control of the opening and closing of the optimizing loop
7 Action law	8 Switching law
9 Sign inverter	10 Switch
11 Integrator	12 Memory
13 Switch	14 Excitation control

(2) *Constant action law*. The behaviour of the system in the case of an action law of constant amplitude has confirmed the practical validity of the theoretical study, despite the differences between the real system and the ideal system studied, especially: the presence of noise, the non-parabolic nature of the static curve, and a more complicated dynamic system.

(3) *Proportional action law*. The behaviour of the first and second derivatives [Figures 14 (c)] during the search for the optimum shows that an action law dependent upon the amplitude of the second derivative allows the amplitude of the oscillations to be reduced, and at the same time reduces the settling time. The effect of disturbances is hence also much reduced. The action law given by the element 7 has the form shown in Figure 19 (a), and the adjustment of the parameters  $k$ ,  $K$ ,  $S_1$  and  $S_2$  allows the best action law to be chosen for each case.

(4) *Average optimum value*. The theoretical study showed, on the hypothesis of a parabolic approximation, the symmetry of the oscillations of type NN around the optimum. This property is very useful since it becomes possible to use these oscillations to derive the average value of the optimum and to impose this value as a control value on the acting variable  $x$ . The oscillations are then eliminated [Figure 18 (b)]. The presence of internal or external disturbances, however, necessitates the elaboration in a systematic manner of new optimum values of the acting variable by closing the control loop. This function is produced by the logic elements in the computer.

**Logic elements**—The logic elements perform the following functions (Figure 17).

(1) *Sign of the action law*. The sign of the action law ( $dx/dt$ ) 9 must be produced according to the proposed switching law. This switching law requires a sequential circuit 8 which does not pose any special problem. This circuit has been produced using semiconductors.

(2) *Memorization of the average optimum value*. The average optimum value is stored in an analogue memory (12) until the production of the next average optimum. The acting variable  $x$  can be connected by means of the static switch (13) either to the integrator (11) comprising the last stage of the optimizing computer which produces  $x$  from  $dx/dt$  (in this case the optimization loop is closed) or to the memory (12) after a given number of switchings. In this case the optimization loop is open; the acting variable  $x$  is regulated to its average optimum value.

(3) *Control of the opening and closing of the optimization loop*. The element (10) which opens and closes the optimization loop is controlled by the logic circuit (6) in order to produce the following functions.

(a) Opening of the control loop (connection of  $x$  to the analogue memory) when the average optimum value is reached.

(b) Closing of the control loop when the second derivative exceeds a predetermined value, i.e. in the presence of sufficiently rapid disturbances.

(c) Closing of the control loop at predetermined time intervals in order to compensate for disturbances which are too slow for their second derivatives to exceed the fixed threshold. Figures 19 (b) and 18 (b) illustrate these different functions:

The approach to the optimum (ABC), the production of the average optimum value (BC), the opening of the control loop and the connection of the acting variable to this average optimum value (point C), and the trend of  $h(t)$  towards its corresponding optimum value (CD). The disturbance appearing at point D [Figure 19 (b)] automatically causes the closing of the control loop and starts a new search cycle (DEFG). Finally, several search cycles from different initial conditions are indicated; compare Figures 20 and 21 with Figure 15 (c), (d) and (e).

It can be finally stated that, according to this conception, the optimizing computer produces the optimum control value, imposes it on the acting variable  $x$ , and modifies it in order to take account of fast or slow disturbances which tend either to displace the system from its optimum working point or to change the optimum value itself.

#### Conclusions

The interesting points of the proposed method, the problems encountered and the directions of future development are now considered.



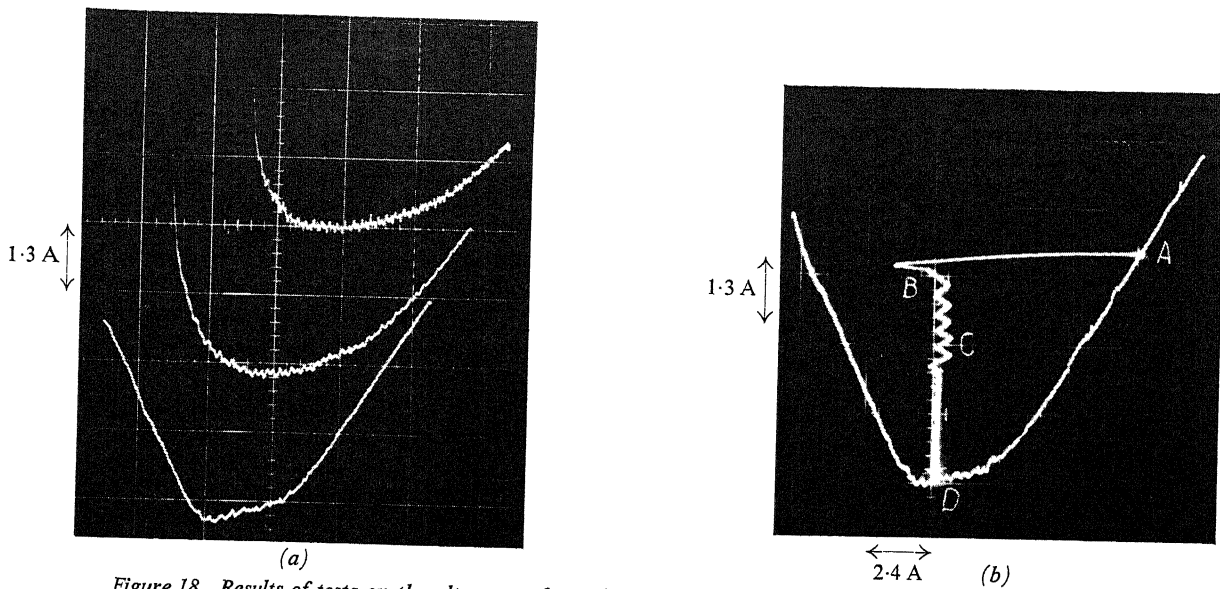


Figure 18. Results of tests on the alternator for different values of the parameters and several initial conditions

(a)  $kW$ : 1, 6/3, 4/5, 2

(b)  $K\tau_{\min} = 1$   
 $K\tau_{\max} = 6$   
 $W = 0.04$

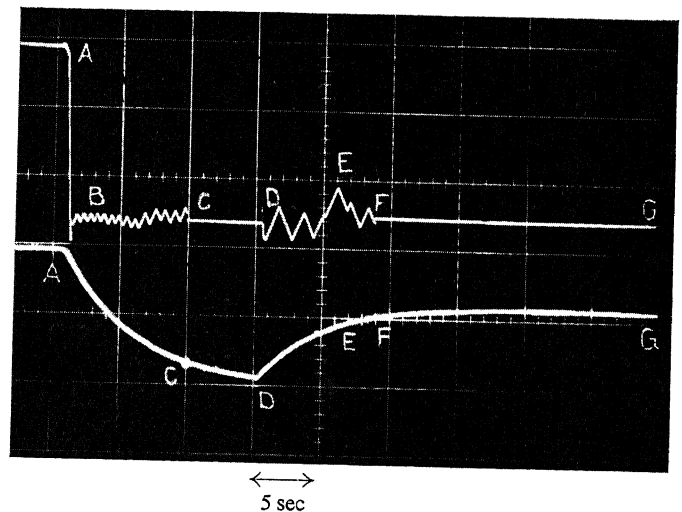
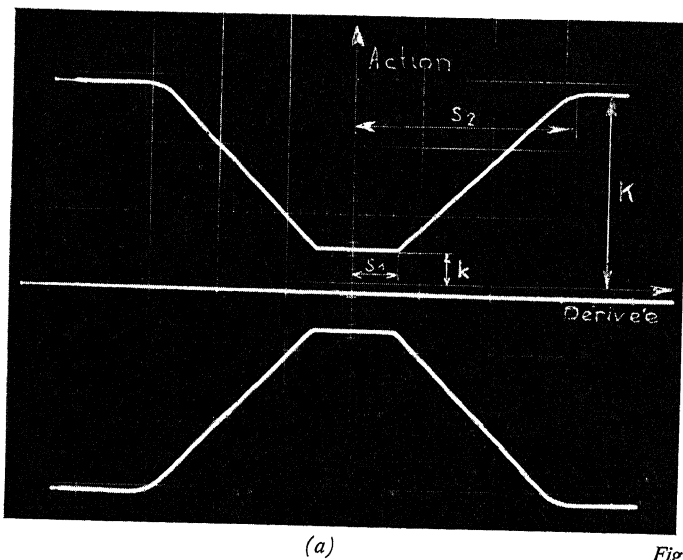


Figure 19

(b)  $K\tau_{\min} = 1$   $K\tau_{\max} = 6$   $W = 0.04$

From a general point of view, methods of automatically searching for the optimum working point have the advantage that they can be applied without necessarily knowing a great deal about the dynamic behaviour of the system, which is necessary in methods which require a mathematical model of the system.

So far as the method itself is concerned, the choice of the sign of a quantity dependent upon the second derivative as a switching criterion has allowed the use of forced switching periods which are very small compared to the system time constant. This is a great advantage since the range of disturbances which the system can tolerate is proportionately increased.

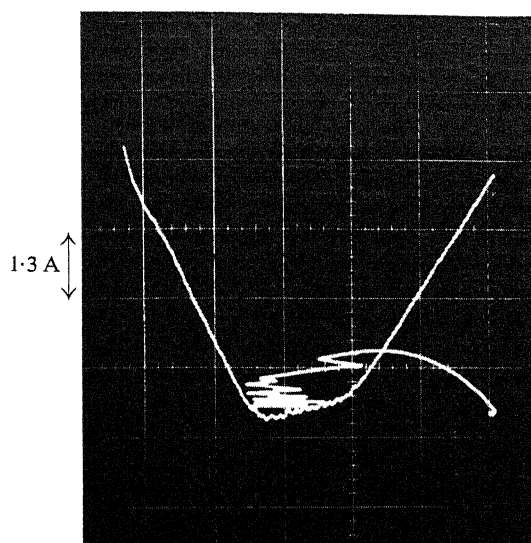
Also the fact of being able to reduce the forced switching time allows an increase in gain, and hence a reduction in settling time,

for a given amplitude of oscillations. The settling time can be said to be of the order of the system time constant.

Notice also the inherent stability of this method of optimization. An unwanted spurious switching, which places the state of the system on a divergent trajectory, is corrected by a forced switching an instant  $W$  later. The same can be said for the appearance of very fast disturbances.

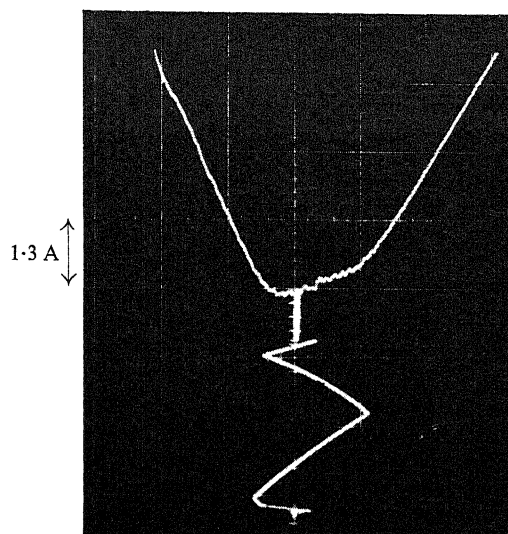
One of the main problems of the proposed method is the production of the first and second derivatives; whatever method is envisaged, analogue, sampling, etc., there remains the problem of separating the useful signal from the noise. This separation becomes easier as the noise and signal have different frequency spectra. Notice that the problem is much simplified if only the sign of the second derivative is required.



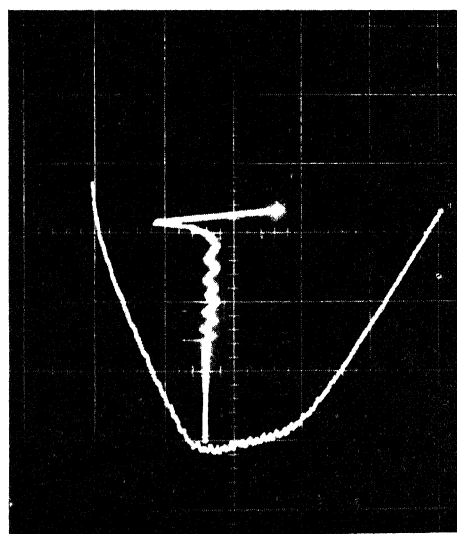


(a)  $K\tau_{\min} = 1$   
 $K\tau_{\max} = 6$   
 $W = 0.04$

Figure 20

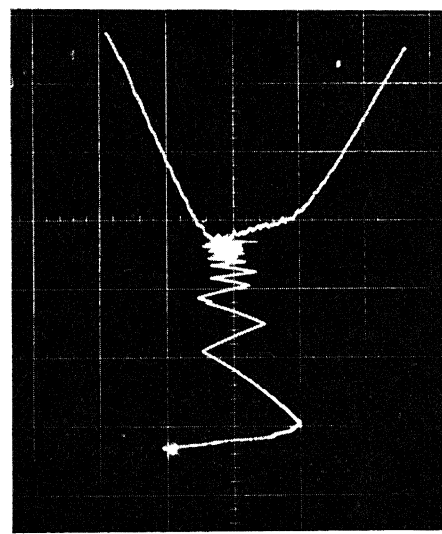


(b)  $K\tau_{\min} = 1$   
 $K\tau_{\max} = 6$   
 $W = 0.04$



(a)  $K\tau_{\min} = 1$   
 $K\tau_{\max} = 6$   
 $W = 0.04$

Figure 21



(b)  $K\tau_{\min} = 1$   
 $K\tau_{\max} = 6$   
 $W = 0.04$

Preliminary tests carried out on the analogue computer showed that this method could be extended to systems with several variables and of higher order. Finally, the proposed method is in the process of application to a chemical reactor, working in the gaseous phase, in order to obtain at all times the maximum efficiency.

This study was carried out at the Battelle Memorial Institute, Geneva, to which the authors extend their thanks. They would also like to thank Mr. R. Bertoldi for his invaluable collaboration.

# Nomenclature

- $f$  Figure of merit ( $\alpha x^2$ )
- $G$  Dimensionless value of  $g$  ( $d^2H/dT^2 - \Delta$ )
- $g$  Switching function ( $d^2h/dt^2 - \delta$  ( $\text{sec}^{-2}$ ))
- $H$  Dimensionless value of  $h$  ( $h/2 \propto K^2\tau^2$ )
- $h$  Dynamic value of figure of merit
- $K$  Constant amplitude of the derivative  $dx/dt$  ( $\text{sec}^{-1}$ )
- $k$  Minimum amplitude of the action law ( $\text{sec}^{-1}$ )
- $p$  Laplace operator ( $\text{sec}^{-1}$ )
- $p_b$  Disturbance

- $R$  Dimensionless value of the dead time ( $\tau_r/\tau$ )  
 $T$  Dimensionless time ( $t/\tau$ )  
 $t$  Time (sec)  
 $W$  Dimensionless switching period ( $\tau_w/\tau$ )  
 $X$  Dimensionless value of  $x$  ( $x/K\tau$ )  
 $x$  Acting variable  
 $\alpha$  Coefficient of the parabola law  
 $\Delta$  Dimensionless value of threshold ( $\delta/2 \propto K^2$ )  
 $\delta$  Switching threshold ( $\text{sec}^{-2}$ )  
 $\varepsilon = \pm 1$   
 $\lambda$  Integration constant  
 $\mu$  Coefficient  $(1 - \Delta) e^{-R}$   
 $\tau$  System time constant (sec)  
 $\tau_e$  Time constant of the acting element (sec)  
 $\tau_r$  Dead time (sec)

## References

- <sup>1</sup> DRAPER, C. S. and LI, J. T. Principles of optimalizing control. *Amer. Soc. mech. Engrs* (1951)
- <sup>2</sup> KAZAKEVICH, V. V. The process of optimizing control of objects with inertia in the presence of disturbances. 1961. *Automatic and Remote Control*. London; Butterworths. Extreme control systems and their improvement. *Automatic control and computer engineering* (1961)
- <sup>3</sup> PUTSILLO, V. P. The principles of construction of one class of extremal control systems for automatized circuits in production processes. 1961. *Automatic and Remote Control*. London; Butterworths
- <sup>4</sup> TAYLOR, W. K. A experimental control system with continuous automatic optimization. 1961. *Automatic and Remote Control*. London; Butterworths
- <sup>5</sup> DOUCE, J. L. and KING, R. E. A self optimizing non-linear control system. *J. Instn. elect. Engrs*. July (1961)
- <sup>6</sup> BOX, G. E. P. and CHANMUGAM, J. Adaptive optimization of continuous processes, *I.E.C. fundamentals*, February (1962)
- <sup>7</sup> MOROSANOV Extremum control methods. *Automat. Telemek.* No. 11 (1957)
- <sup>8</sup> FRAIT, J. S. and ECKMAN, P. Optimizing control of single input extremum systems. *J. Basic Engng.* March (1962)

## DISCUSSION

M. HAMZA, ETH, Zurich, Switzerland

Extremum seeking regulators have a wide and important field of application and several methods have been suggested for their design<sup>1-3</sup>. The system suggested by Perret and Rouxel is novel, especially their method of employing the second derivative, and they are to be highly congratulated for their excellent work. A technique will be presented which does not possess some of the disadvantages of the system suggested in this paper, namely:

- (1) Its performance is dependent on initial conditions (*Figure A*).
- (2) Time is required to build the average, *Figure 20(b)* of the paper.
- (3) If the non-linearity is not symmetrical, the method of building the average will not lead to extremum.

The following technique was derived for higher-order systems having various types of non-linearities and for systems having  $n$  variables. A paper describing this technique in full, with several illustrations, will be published soon. The method is based on determining the slope of the non-linear function by measuring the discontinuity in an appropriate derivative of the system's output when the input is (for convenience) a step. In some respects, this method is similar to the sinusoidal perturbation technique<sup>3</sup> and possesses its advantages. As an illustration, with reference to *Figure A* and assuming no delay to be present and that the integrator is not a perfect one, this may be

given by  $1/\tau_1 p + 1$ , the discontinuity in the second derivative of the system's output with respect to time when the system input is a step of magnitude  $A$  may be easily seen to be

$$\frac{d^2 h(0^*)}{dt^2} = \frac{A}{p} \frac{1}{(\tau_1 p + 1)} K^* \frac{1}{\tau p + 1} p^2 \cdot p \Big|_{p \rightarrow \infty} \quad (1)$$

$$= \frac{AK^*}{\tau_1 \tau} \quad (2)$$

where

$$K^* = \frac{d}{dx} [f(x)] \Big|_{x=x_0} \quad (3)$$

$x_0$  refers to  $x$  at  $t = 0$ . Note that

$$\text{sign} \frac{d^2 h(0^*)}{dt^2} = \text{sign} \frac{AK^*}{\tau_1 \tau} \quad (4)$$

$$\left| \frac{d^2 h}{dt^2} \right| = \left| \frac{AK^*}{\tau_1 \tau} \right| \quad (5)$$

If  $f = \alpha x^2$ , then

$$K^* = 2\alpha x \quad (6)$$

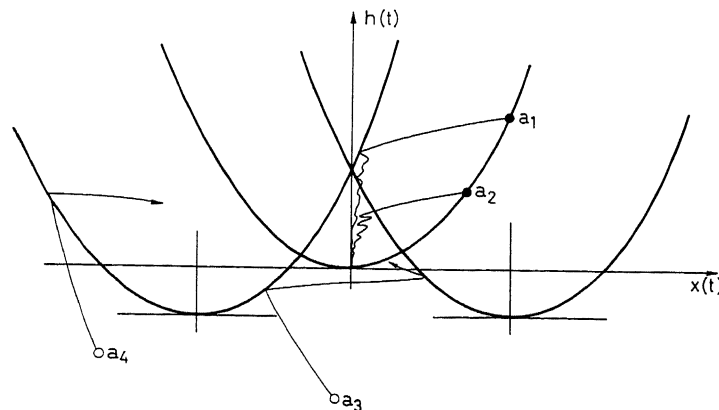


Figure A

and

$$\frac{d^2 h(0^*)}{dt^2} = \frac{2 A \alpha}{\tau_1 \tau} x \quad (7)$$

Thus a simple logic circuit governed by eqn (4) or by eqns (4) and (5)—where eqns (4) and (5) are first used to estimate the position of the extremum, since the discontinuity is proportional to  $x$ , and once near the extremum eqn (4) alone—may be used to govern the behaviour of the controller. Near the extremum the oscillations may be reduced using prediction linearization and other techniques as suggested in<sup>1,2</sup>. If a delay is present it may be determined by measuring the difference in time between the application of the input step—or disturbance—and the occurrence of the discontinuity in the system output. Thus the effect of the delay may be taken into account.

In the original version of Perret and Rouxel's paper, which was written in French, they showed that the second derivative suffered a discontinuity proportional to  $X$ , and Nour Eldin (in a colloquium) pointed out that for the case studied by the latter authors the discontinuity may be used to identify the extremum. He obtained this result independent of the author and his method was based on a phase plane study.

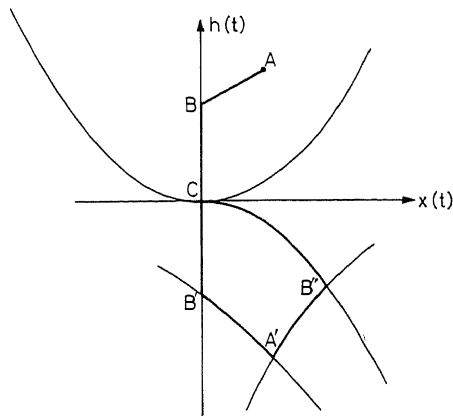


Figure B

I would like to point out that the results in Figure 15 (a) and (b) might be slightly misleading since  $K\tau_{\min}$  and  $K\tau_{\max}$  in both cases are not the same. Further, how could the first switching in Figure 15 (e) be explained? It does not occur on a parabola and it must be of the  $N$  type. Finally, it is worth noting that although the shortest time to reach the extremum from  $A$ , Figure B, is  $ABC$ , the shortest from point  $A'$  is not  $A'B'C'$  but could be  $A'B''C$ , as may be readily proved.

#### References

- <sup>1</sup> DRAPER, C. S. and LI, J. T. Principles of optimizing control. *Amer. Soc. Mech. Engrs* (1951)
- <sup>2</sup> FELDBAUM, A. A. *Computers in Automatic Control Systems*. 1959. Moscow; Gifml
- <sup>3</sup> WESTCOTT, J. H. *An Exposition of Adaptive Control*. 1962. Oxford; Pergamon Press

H. NOUR ELDIN, *Swiss Federal Institute of Technology, ETH, Zurich, Switzerland*

The paper presented by Perret and Rouxel is interesting as it uses the second derivative for control law. Using the second derivative has its drawbacks and I shall demonstrate an inherent property in this hill climbing problem.

The dynamic characteristics of the system can be shown in the two planes  $U^+$  and  $U^-$ . At point  $a^+$  the control variable  $U$  is negative and crosses the zero acceleration point. If the control variable is switched to  $U^+$  at this point, the point  $a^+$  will be point  $a^-$  in the  $U^+$  plane. At

this point its acceleration is positive. There will be a jump in the acceleration at the switching point  $a$ . The equation of this jump is:

$$\text{jump amplitude} = f(\alpha_1 T) x$$

where  $\alpha$  is the order of the parabola and  $T$  the time constant of the system. If the plant characteristic does not vary during this time the jump amplitude will be directly proportional to the distance from the minimum. The direction of the jump depends on the transition direction of the control variable  $U$ . These jumps are shown in Figure A. If we use this jump property for detection of the distance from the minimum, we choose a certain perturbation sequence, say  $U^-$  to  $U^+$ . Measuring the jump in the second derivative will give the distance from the minimum. For varying plant characteristics one can identify the time constant and calculate the distance from the minimum.

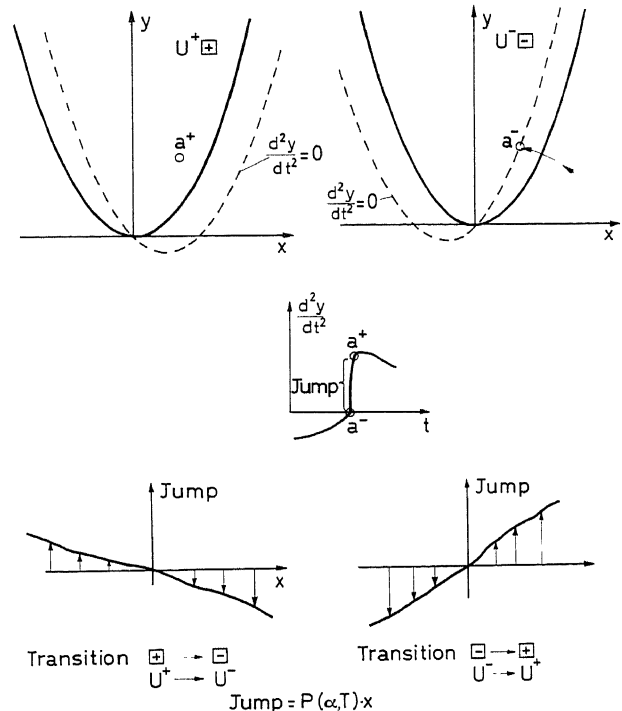


Figure A

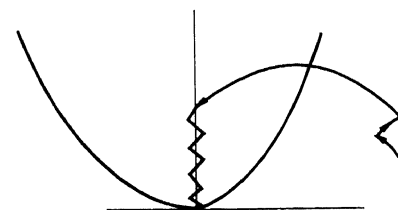
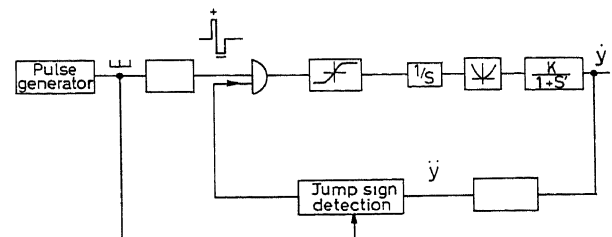


Figure B

The jump property can be used for one-dimensional and also for multi-dimensional systems and it can be extended to multi-valley problems. It works also in the presence of noise (by averaging before and after transition). In the presence of delay time the system can work

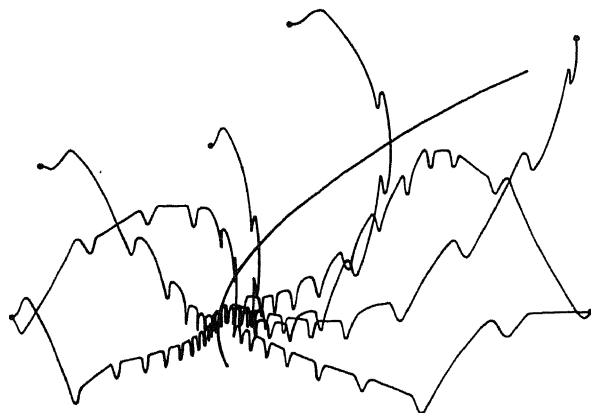


Figure C

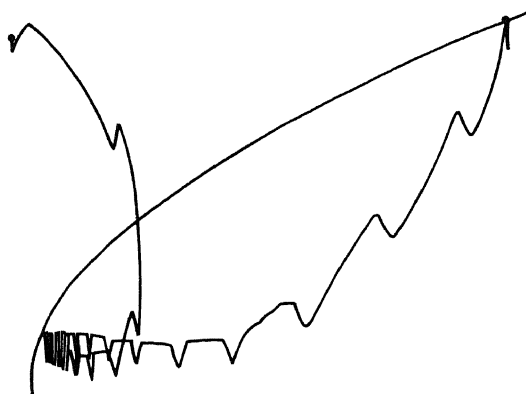


Figure D

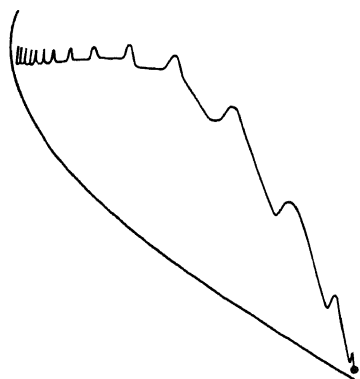


Figure E

with modification for delay estimation. For a plant of higher order, Mr. Hamza at the Swiss Federal Institute of Technology, working independently, has proved the jump property for the higher-order system. One can build a hill climbing system in a very simple manner using a logic circuit which determines the sign of the jump. This system does not require any plant identification. It measures the second derivative before and after the transition and determines whether the plant is at the right or at the left from the minimum. The block diagram and the results are shown in Figures B to E. The system goes to the minimum with good accuracy. May I remark that reaching the minimum point in the minimum time can be achieved by switching off the system at  $X = 0$  (or oscillations around  $X = 0$ ).

R. PERRET AND R. ROUXEL, *in reply*

The improvement prepared by Hamza and by Nour Eldin presents some advantages for first-order systems, especially for dynamic initial conditions located in zone 1 of Figure 6 of our paper. Nevertheless, we would like to make the following remarks:

(1) For  $n$ th order systems, this method implies the elaboration of the  $(n + 1)$ th time derivative which is a severe practical limitation.

(2) This method needs a periodic test signal superimposed on the actuating one. The dead time inherent to physical systems implies that between every test input and test output signal a time delay at least equal to the dead time is required which seems to be difficult to compensate. The method proposed by the authors avoids these disadvantages. The last two questions of M. Hamza can be answered as follows.

(1) In Figure 15(b) of the paper, the maximum value of the gain is higher than the gain related to Figure 15(a). For a given action law, the amplitude of the oscillations usually increases with the gain. In this specific case [Figure 15(b)] with a gain proportional to the second derivative, the amplitude of oscillations, instead of increasing, is reduced. This demonstrates the advantage of using such an action law.

(2) In Figure 15(e), a false commutation, due to the noise, effectively appeared at the beginning of the evolution. In spite of this fact, the system converges towards the optimum that illustrates the self-stability inherent to the method.

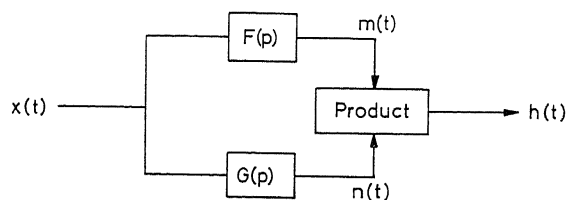


Figure A

Finally, we would like to point out that the study of different applications of this method, specially in the field of chemical processes, has led us to consider systems with a different structure as it is shown in Figure A. In this case the index of performance  $h(t)$  is issued from the product of two signals  $m(t)$  and  $n(t)$  obtained as  $F(p)$  and  $G(p)$ , the input of two transfer functions. The study of this category of systems is being made at present.

# Dual Control Theory Problems

A. A. FELDBAUM

## Summary

Control in closed systems, with incomplete *a priori* information about the plant, should, in the general case, include two types of action on the plant: (a) tentative actions, 'experiments', with the controlled plant, whose problem consists in studying it; (b) directing actions, whose task consists in adjusting the plant to the required operating conditions. The controlling action in optimum control, should, generally speaking, realize both of the above-mentioned tasks simultaneously, and because of this, has a dual character. Dual control problems are considered in this paper. The problem is set up, and its general solution given, for discrete-time systems, when a change in the plant's characteristics represents a discrete Markov process. Various examples are considered.

## Sommaire

Le réglage de systèmes en boucle fermée, dans lesquels les informations *a priori* sur l'installation à régler sont incomplètes, doit comprendre, d'une manière générale, deux types d'actions exercées sur cette installation:

(a) des «expérimentations» en vue de connaître le comportement de cette dernière; (b) des actions de réglage en vue d'imposer à l'installation un comportement qui satisfasse certains critères.

Dans le réglage optimal, l'action de commande doit, en principe, réaliser ces deux types d'actions simultanément. Elle possède, de ce fait, un caractère dual. Ce rapport a pour objet les problèmes de réglage à caractère dual. Il pose le problème et en donne la solution générale pour les systèmes discontinus dans le temps, dans lesquels les changements des caractéristiques de l'installation constituent une chaîne discontinue de Markov. Divers exemples y sont passés en revue.

## Zusammenfassung

Die Regelung in geschlossenen Kreisen mit unvollständiger *a-priori* Information über die Strecke sollte im allgemeinen Fall zwei Arten von Eingriffen auf die Strecke umfassen:

a) studierende Einwirkungen auf die Regelstrecke („Experimente“) und  
b) steuernde Eingriffe, die die Regelstrecke in den geforderten Betriebszustand versetzen sollen.

Bei der Optimalregelung sollen im allgemeinen diese beiden Funktionen gleichzeitig erfüllt werden, die Regelwirkung bekommt dadurch dualen Charakter. In diesem Beitrag werden duale Regelprobleme behandelt. Das Problem wird definiert und eine allgemeine Lösung wird für zeit-diskrete Systeme angegeben, sofern eine Änderung der Regelstreckenmerkmale einen diskreten Markoff-Prozess darstellt. Einige Beispiele werden angeführt.

## Introduction

The controlling device in an automatic system solves two problems that are closely interrelated, but which differ in character. In the first place, on the basis of information that is fed into it, it clarifies the properties and state of the controlled plant. In the second place, on the basis of the properties discovered in the plant, it determines which steps have to be taken for successful control. The first task is that of studying the

plant; the second task is that of adjusting the plant to the required operating conditions. In the simplest types of systems, the solution of one of these problems may be absent or have a primitive form. In complex cases, the controlling device should solve both indicated problems. Below are considered problems involved in the design of optimal devices that solve both problems simultaneously.

Optimal systems may be divided into three types: (a) optimal systems having complete or the maximum information possible about the controlled plant; (b) optimal systems having incomplete information about the plant, and with an independent or passive accumulation of it in the control process; (c) optimal systems having incomplete information about the plant, and with an active accumulation of it in the control process.

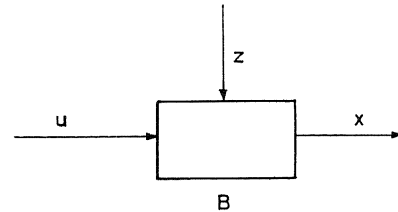


Figure 1

In order to clarify this classification, it is necessary to determine what is meant by information regarding the controlled plant. In Figure 1, the controlled plant *B* is shown in the shape of a rectangle; in it, *x* is the output quantity, *u* is the controlling action and *z* is a non-controlled disturbance. If the plant has several inputs and outputs, then the quantities *x*, *u*, *z* must be considered as vectors.

$$\text{The dependence} \quad x = F(u, z) \quad (1)$$

may assume the form of a relationship between the values of *x*, *u*, *z* at the same moment of time either in the form of a differential or some other type of equation. In the general case, *F* is an operator.

Information regarding the plant is gathered from the following elements: (a) information about the plant operator *F*; (b) information about the disturbance *z*, which acts on the plant; (c) information about the state of the plant—for example, about the coordinates of the point that represents the state of the plant in the phase space; and (d) information regarding the control target.

The last-mentioned information element should indicate the ideal that is to be attained, and also the 'price' of a deviation from this ideal. For this reason, the control goal may be conveniently presented in the form of a requirement for the minimization of some functional *Q*, which depends on the character of the *x*, *u*, *z* processes, and also on some externally assigned process *x\**, the reference quantity.

Below is a statement limited by conditions of the type:

$$Q(x, x^*) = \min \quad (2)$$

Let  $Q$  be called the optimization criterion. Set the requirement, for example, that the ideal process  $x$  be identical with  $x^*$ , and that the 'cost' of the deviation from the ideal be expressed by the formula:

$$Q = c \int_0^T (x - x^*)^2 dt \quad (3)$$

where  $c$  and  $T$  are constants. Expression (3) is an example of the optimization criterion. A system called optimal in which the minimum of the criterion  $Q$  is realized, while fulfilling the additional conditions that characterize the problem—for example, that  $u$  and  $x$  belong to some admissible domains  $\Omega(u)$  and  $\Omega(x)$ , respectively.

Complete information regarding some arbitrary dependence implies absolutely accurate knowledge its. For example, complete information about some time function,  $f(t)$ , denotes that its values are known for any arbitrary values of  $t$ . It is considered below, that complete information is available regarding the operator  $F$ , and also regarding the optimization criterion  $Q$ . All that is unknown and unforeseen in the plant is attributed to the disturbance  $z$ , and what is unknown in the control goal—to the reference quantity  $x^*$ .

Theories of optimal systems, with complete information about the plant, were developed in a number of papers<sup>1-4</sup>. In the theory regarding systems with an independent accumulation of information about the plant, consideration was given, for the main part, to open-loop systems. Here statistical study methods were introduced<sup>1, 5, 7</sup>.

Problems have been considered<sup>8, 9</sup> that pertain to the third type of optimal system theories. The theory involving systems of the third type contains characteristics that are common both to theories of the first as well as to those of the second type, and which therefore unite them to a certain extent. However, the third type of theory is also characterized by its own specific features.

The block diagram, previously studied<sup>8, 9</sup>, is shown in Figure 2. A closed-loop system is considered, in which the operator of the plant  $B$  and the optimization criterion  $Q$  have been assigned. Plant  $B$  is acted on by a random disturbance  $z$ , which cannot be measured directly. The controlling action  $u$  is admitted from the controlling device  $A$  to the plant  $B$ , through the connecting channel  $G$ , where it is mixed with the random noise  $g$ . For this reason, the action  $v$ , at the input of the plant  $B$ , is not equal, generally speaking, to the quantity  $u$ . Further, information regarding the plant's state  $x$  goes through the connecting channel  $H$ , where it is mixed with the random noise  $h$ . The output  $y$  of the connecting channel  $H$  is admitted to the input of the controlling device  $A$ . Reference quantity  $x^*$  is also admitted to the input of device  $A$  through channel  $H^*$ , with noise  $h^*$ .

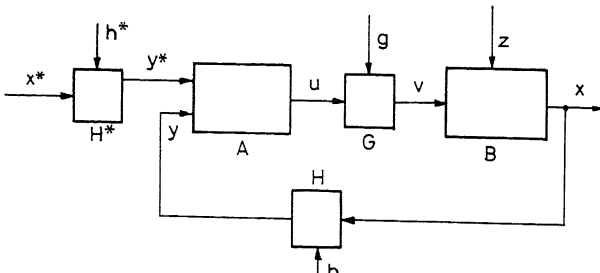


Figure 2

In the circuit shown in Figure 2, processes are possible that have not been considered in the theories of the first two types. A study of disturbance  $z$ , i.e., essentially, of the changing characteristics of plant  $B$ , may be carried out in the circuit of Figure 2, not by means of passive observation, but through an active method, by means of rational experiments. The plant is 'felt', as it were, by the  $u$  actions, which have a perceptive character, and the results  $y$  of these actions are analysed by the  $A$  device. The purpose of these actions is to promote a more rapid and accurate study of the plant's characteristics; this can develop a better principle for controlling it.

However, the controlling action is necessary not only for the purpose of study, but also for directing and adjusting the plant to the required operating conditions. Consequently, in the circuit of Figure 2, the controlling actions should have a dual character; to a certain degree they should be studying actions, but they should also be directing actions. That is why the theory underlying systems of this type is called a dual-control theory.

It is precisely this duality of control that constitutes the physical fact distinguishing the third type of optimal systems from the first two. In the first one, dual control is not necessary, in so far as the controlling device, without it, possesses complete (or maximally possible) information about the plant. In the second type of system, dual control is impossible, because the information is accumulated by means of observation alone, and the rate of its accumulation does not depend at all on the strategy of the controlling device.

### Setting Up the Problem

Theoretical problems are considered below only for systems that are time discrete. All the quantities indicated in the circuit in Figure 2 are considered only for discrete time moments,  $t = 0, 1, 2, \dots, n$ , where  $n$  has been fixed. The value of any arbitrary quantity for an  $s$  discrete time has been provided with an index  $s$ —for example,  $x_s^*$ ,  $x_s$ ,  $y_s$ , etc. The transmission lines of all quantities are assumed to be single-channelled, and the plant  $B$  is considered to have no memory. Therefore, its equation may be written thus:

$$x_s = F(v_s, z_s) \quad (4)$$

A generalization of the conclusion set forth below, for more complex plant cases, with several inputs and outputs, and also for plant having a memory, may be carried out in the same way as was done previously<sup>8</sup>.

Let  $h_s^*$ ,  $h_s$  and  $g_s$  represent a series of independent, random quantities with invariable distribution densities:  $P(h_s^*)$ ,  $P(h_s)$ ,  $P(g_s)$ . Further, let:

$$z_s = z_s(s, \bar{\mu}_s) \quad (5)$$

and

$$x_s^* = x_s^*(s, \bar{\lambda}_s) \quad (6)$$

where the vectors  $\bar{\mu}_s$  and  $\bar{\lambda}_s$ , in contrast with refs. 8 and 9, are not random quantities, but discrete vector Markov random processes. In other words, the  $\bar{\mu}_s$  and  $\bar{\lambda}_s$  are vectors:

$$\bar{\mu}_s = (\mu_s^1, \dots, \mu_s^m) \quad (7)$$

and

$$\bar{\lambda}_s = (\lambda_s^1, \dots, \lambda_s^l) \quad (8)$$

$\mu_s^i$  and  $\lambda_s^j$ , in the general case, are mutually interrelated, scalar Markov discrete processes. The same holds for  $\lambda_s^i$  and  $\lambda_s^j$ .

However, the vectors  $\bar{\mu}_s$ ,  $\bar{\lambda}_s$ , and also, noises  $h_s^*$ ,  $h_s$ ,  $g_s$  are considered independent.

Consider the Markov process characteristics  $\bar{\mu}_s$  and  $\bar{\lambda}_s$  as having been given. This implies that one has been given, as the initial probability densities,  $P_0(\bar{\mu}_0)$  and  $P_0(\bar{\lambda}_0)$  where  $t = 0$ , as well as the transient probability densities  $P(\bar{\mu}_{i+1}/\bar{\mu}_i)$  and  $P(\bar{\lambda}_{i+1}/\bar{\lambda}_i)$ .

The methods for combining the signal and the noise in blocks  $H^*$ ,  $H$  and  $G$  are considered as known and invariable, and the blocks themselves as having no memory. Therefore:

$$v_s = v_s(u_s, g_s); y_s^* = y_s^*(h_s^*, x_s^*); y_s = y_s(h_s, x_s) \quad (9)$$

The control goal is determined in the following manner: let the specific loss function (the 'cost' of deviation from the ideal), which corresponds to the  $s$  time, have the form:

$$W_s = W_s(s, x_s, x_s^*) \quad (10)$$

Further, let the overall loss function  $W$ , for the entire time period  $n$ , be equal to the sum of the specific loss functions:

$$W = \sum_{s=0}^{s=n} W_s(s, x_s, x_s^*) \quad (11)$$

A system is called optimal for which the average risk  $R$  (mathematical expectation  $M$  of the magnitude  $W$ ) is minimal. The risk magnitude is expressed by the formula:

$$R = M\{W\} = M\left\{\sum_{s=0}^{s=n} W_s(s, x_s, x_s^*)\right\} = \sum_{s=0}^{s=n} M\{W_s\} = \sum_{s=0}^{s=n} R_s \quad (12)$$

The expression  $R_s = M(W_s)$  is called the specific risk in the  $s$  cycle. The quantity  $R$  plays the part, here, of the optimization criterion  $Q$ .

Introduce the time vectors ( $0 \leq s \leq n$ ):

$$\left. \begin{aligned} \vec{u}_s &= (u_0, u_1, \dots, u_s); & \vec{x}_s^* &= (x_0^*, x_1^*, \dots, x_s^*) \\ \vec{v}_s &= (v_0, v_1, \dots, v_s); & \vec{y}_s &= (y_0, y_1, \dots, y_s) \\ \vec{x}_s &= (x_0, x_1, \dots, x_s); & \vec{y}_s^* &= (y_0^*, y_1^*, \dots, y_s^*) \end{aligned} \right\} \quad (13)$$

and the matrices of vector  $\bar{\mu}_s$ , vector  $\bar{\lambda}_s$ , which are made up of the vector columns of  $\bar{\mu}_s$ ,  $\bar{\lambda}_s$ :

$$\bar{\mu}_s = (\bar{\mu}_0, \bar{\mu}_1, \dots, \bar{\mu}_s); \quad \bar{\lambda}_s = (\bar{\lambda}_0, \bar{\lambda}_1, \dots, \bar{\lambda}_s) \quad (14)$$

Consider that the control device, in the general case, possesses a memory. In addition to this, assume, for general purposes, that the algorithm of this device is a random one. The term 'random strategy' is also employed. This implies that the value  $u_s$  is a random function of the quantities  $y_i$  which were admitted to the input of device  $A$  during the preceding moments of  $u_i$  ( $i < s$ ), and also of the values  $y_j^*$  ( $j \leq s$ ). It is required to find the optimal probability densities:

$$P_s(u_s) = \Gamma_s(u_s | \vec{u}_{s-1}, \vec{y}_{s-1}, \vec{y}_s^*)_{(0 \leq s \leq n)} \quad (15)$$

The problem consists in finding such a series of functions  $\Gamma_s$ , in the case of which the average risk  $R$  will be minimal. Inasmuch as  $\Gamma_s$  is the probability density, therefore:

$$\int_{\Omega(u_s)} \Gamma_s(u_s) d\Omega(u_s) = 1 \quad (16)$$

where  $\Omega(u_s)$  designates the region for the magnitude changes  $u_s$ , and  $d\Omega(u_s)$  represents its infinitely small element. And thus it must be found that the optimal functions  $\Gamma_s \geq 0$ , which are limited by condition (16).

### Derivation of the Basic Formula

First write the formula for a conditional, specific risk,  $r_s$ , understanding by this a risk in the  $s$  cycle, with a fixed 'pre-history' of the control device inputs, i.e., with fixed values for  $\vec{y}_s^*$ ,  $\vec{y}_{s-1}$ ,  $\vec{u}_{s-1}$ :

$$\begin{aligned} r_s &= M\{W_s | \vec{y}_s^*, \vec{u}_{s-1}, \vec{y}_{s-1}\} \\ &= \int_{\Omega(\bar{\lambda}_s, x_s)} W_s[s, x_s, x_s^*(s, \bar{\lambda}_s)] \cdot P(\bar{\lambda}_s, x_s | \vec{y}_s^*, \vec{u}_{s-1}, \vec{y}_{s-1}) d\Omega(\bar{\lambda}_s, x_s) \end{aligned} \quad (17)$$

Here  $\Omega(\bar{\lambda}_s, x_s)$  is the domain of changes for  $\bar{\lambda}_s$  and  $x_s$ , and  $d\Omega(\bar{\lambda}_s, x_s)$  is its infinitely small element;  $P(\bar{\lambda}_s, x_s | \vec{y}_s^*, \vec{u}_{s-1}, \vec{y}_{s-1})$  is the conditional, common probability density of the  $\bar{\lambda}_s$  and  $x_s$ , with fixed vectors  $\vec{y}_s^*$ ,  $\vec{u}_{s-1}$ ,  $\vec{y}_{s-1}$ . In conformity with a well-known theorem of the theory of probabilities, an equality exists:

$$\begin{aligned} &P(\bar{\lambda}_s, x_s | \vec{y}_s^*, \vec{u}_{s-1}, \vec{y}_{s-1}) \\ &= P(\bar{\lambda}_s | \vec{y}_s^*, \vec{u}_{s-1}, \vec{y}_{s-1}) \cdot P(x_s | \vec{y}_s^*, \vec{u}_{s-1}, \vec{y}_{s-1}, \bar{\lambda}_s) \\ &= P(\bar{\lambda}_s | \vec{y}_s^*) \cdot P(x_s | \vec{y}_s^*, \vec{u}_{s-1}, \vec{y}_s) \end{aligned} \quad (18)$$

The last transformation is accurate, because the probability density of  $\bar{\lambda}_s$ , with a fixed vector  $\vec{y}_s^*$ , will not change if vector  $\vec{u}_{s-1}$ , vector  $\vec{y}_{s-1}$  are also fixed (see Figure 2). Further, the probability density of  $x_s$  with a fixed vector  $\vec{y}_s^*$ , will not change, if in addition  $\bar{\lambda}_s$  is fixed. The second multiple (18) is rewritten in an expanded form:

$$\begin{aligned} &P(x_s | \vec{y}_s^*, \vec{u}_{s-1}, \vec{y}_{s-1}) \\ &= \int_{\Omega(\bar{\mu}_s, u_s)} P(x_s | \bar{\mu}_s, u_s) P_s(\bar{\mu}_s) \Gamma_s(u_s | \vec{y}_s^*, \vec{u}_{s-1}, \vec{y}_{s-1}) d\Omega(\bar{\mu}_s, u_s) \end{aligned} \quad (19)$$

where  $\Omega(\bar{\mu}_s, u_s)$  is the domain of changes for  $\bar{\mu}_s$  and  $u_s$ , and  $P_s(\bar{\mu}_s)$  is the *a posteriori* probability density of  $\bar{\mu}_s$  in the  $s$  cycle:

$$\begin{aligned} P_s(\bar{\mu}_s) &= P(\bar{\mu}_s | \vec{u}_{s-1}, \vec{y}_{s-1}, \vec{y}_s^*) \\ &= \int_{\Omega(\bar{\mu}_{s-1})} P(\bar{\mu}_s | \bar{\mu}_{s-1}) P(\bar{\mu}_{s-1} | \vec{y}_{s-1}, \vec{u}_{s-1}, \vec{y}_s^*) d\Omega(\bar{\mu}_{s-1}) \end{aligned} \quad (20)$$

Inasmuch as:

$$\begin{aligned} &P(\bar{\mu}_{s-1} | \vec{y}_{s-1}, \vec{u}_{s-1}, \vec{y}_s^*) \\ &= \int_{\Omega(\bar{\mu}_{s-2})} P(\bar{\mu}_{s-1} | \vec{y}_{s-1}, \vec{u}_{s-1}, \vec{y}_s^*) d\Omega(\bar{\mu}_{s-2}) \end{aligned} \quad (21)$$

where  $\Omega(\bar{\mu}_{s-2})$  is the domain of changes for the matrix  $(\bar{\mu}_{s-2})$ , therefore, it is necessary to find the conditional probability density of  $P(\bar{\mu}_{s-1} | \vec{y}_{s-1}, \vec{u}_{s-1}, \vec{y}_s^*)$ , for the matrix  $\bar{\mu}_{s-1}$ . From the equality:

$$\begin{aligned} &P(\bar{\mu}_{s-1}, \vec{u}_{s-1}, \vec{y}_{s-1} | \vec{y}_s^*) \\ &= P(\vec{u}_{s-1}, \vec{y}_{s-1} | \bar{\mu}_{s-1}, \vec{y}_s^*) \cdot P(\bar{\mu}_{s-1} | \vec{y}_s^*) \\ &= P(\bar{\mu}_{s-1} | \vec{u}_{s-1}, \vec{y}_{s-1}, \vec{y}_s^*) \cdot P(\vec{u}_{s-1}, \vec{y}_{s-1} | \vec{y}_s^*) \end{aligned} \quad (22)$$

one finds (because  $\vec{\mu}_{s-1}$  does not depend on  $\vec{y}_s^*$ ):

$$P_{s-1}(\vec{\mu}_{s-1}) = P(\vec{\mu}_{s-1} | \vec{u}_{s-1}, \vec{y}_{s-1}, \vec{y}_s^*) \\ = \frac{P(\vec{u}_{s+1}, \vec{y}_{s-1} | \vec{\mu}_{s-1}, \vec{y}_s^*) P(\vec{\mu}_{s-1})}{P(\vec{u}_{s-1}, \vec{y}_{s-1} | \vec{y}_s^*)} \quad (23)$$

Here  $P(\vec{u}_{s-1}, \vec{y}_{s-1} | \vec{y}_s^*)$  is the common *a priori* probability density of vectors  $u_{s-1}, y_{s-1}$ , with fixed  $\vec{y}_s^*$ ;  $P(\vec{\mu}_{s-1})$  is the *a priori* probability density of the matrix  $\mu_{s-1}$ , and  $P(\vec{u}_{s-1}, \vec{y}_{s-1} | \vec{\mu}_{s-1}, \vec{y}_s^*)$  is the conditional probability density of vectors  $\vec{u}_{s-1}, \vec{y}_{s-1}$ , with a fixed matrix  $\vec{\mu}_{s-1}$  and vector  $\vec{y}_s^*$  (the likelihood function). In passing ( $s-1$ ) times around the closed loop in Figure 2, it is possible, as in the derivation<sup>8, 9</sup>, to find the expression:

$$P_{s-1}(\vec{\mu}_{s-1}) = \frac{P_0(\mu_0) \prod_{i=1}^{s-1} P(\vec{\mu}_i | \vec{\mu}_{i-1}) \left[ \prod_{i=0}^{s-1} P(y_i | \mu_i, i, u_i) \right] \left[ \prod_{i=0}^{s-1} \Gamma_i \right]}{P(\vec{u}_{s-1}, \vec{y}_{s-1} | \vec{y}_s^*)} \quad (24)$$

The substitution of (24) in (21), and further on, of (21) in (20), (20) in (19) and (19) in (18) will make it possible to determine the second co-factor in (18). Now consider the first co-factor of this expression—the *a posteriori* probability density of  $\vec{\lambda}_s$ :

$$P_s(\vec{\lambda}_s) = P(\vec{\lambda}_s | \vec{y}_s^*) \\ = \int P(\vec{\lambda}_s | \vec{y}_s^*) d\Omega(\vec{\lambda}_{s-1}) \quad (25)$$

Inasmuch as:

$$P(\vec{\lambda}_s, \vec{y}_s^*) = P(\vec{\lambda}_s) \cdot P(\vec{y}_s^* | \vec{\lambda}_s) = P(\vec{\lambda}_s | \vec{y}_s^*) \cdot P(\vec{y}_s^*)$$

therefore:

$$P_s(\vec{\lambda}_s) = P(\vec{\lambda}_s | \vec{y}_s^*) = P(\vec{\lambda}_s) \cdot \frac{P(\vec{y}_s^* | \vec{\lambda}_s)}{P(\vec{y}_s^*)} \quad (26)$$

The *a priori* probability density of the matrix  $\vec{\lambda}_s$  is determined from the formula that is accurate for the Markov process:

$$P(\vec{\lambda}_s) = P(\vec{\lambda}_0, \vec{\lambda}_1, \dots, \vec{\lambda}_s) \\ = P_0(\vec{\lambda}_0) \cdot P(\vec{\lambda}_1 | \vec{\lambda}_0) \cdot P(\vec{\lambda}_2 | \vec{\lambda}_1) \dots P(\vec{\lambda}_s | \vec{\lambda}_{s-1}) \\ = P_0(\vec{\lambda}_0) \prod_{i=1}^s P(\vec{\lambda}_i | \vec{\lambda}_{i-1}) \quad (27)$$

Further, the conditional probability density is:

$$P(\vec{y}_s^* | \vec{\lambda}_s) = P(y_0^* | \lambda_0) \cdot P(y_1^* | \lambda_1) \dots P(y_s^* | \lambda_s) \\ = \prod_{i=0}^s P(y_i^* | \lambda_i) \quad (28)$$

Consequently, from (26), (27) and (28), one obtains:

$$P_s(\vec{\lambda}_s) = \frac{P_0(\vec{\lambda}_0) \prod_{i=1}^s P(\vec{\lambda}_i | \vec{\lambda}_{i-1}) \cdot \prod_{i=0}^s P(y_i^* | \lambda_i)}{P(\vec{y}_s^*)} \quad (29)$$

By substituting (29) in (25), the final formula for  $P_s(\vec{\lambda}_s)$  is found.

Attention is now turned to the principal difference between

formula (24) and (29) for the *a posteriori* probability densities of the matrices  $\vec{\mu}_{s-1}$  and  $\vec{\lambda}_s$ . The accumulation of information regarding the disturbance  $z$  or the vector  $\vec{\mu}_s$ , i.e., essentially, regarding the unexpected manner involving the changing characteristics of the plant, is expressed in the fact that the *a priori* probability density  $P_0(\vec{\mu}_0)$  is replaced in each new cycle by the *a posteriori* densities  $P_s(\vec{\mu}_s)$  [see (20), associated with the expression (23)]. From (23) and (20) it is evident that the function  $P_s(\vec{\mu}_s)$ , and, consequently, also the rate of information accumulation depends on all the preceding strategies  $\Gamma_i$  ( $i < s$ ). In other words, the rate involved in studying the plant depends on how efficiently the experiments were set up with respect to studying this plant, feeding it with the  $u_i$  actions and making analysis of the plant  $y_i$  reactions to these actions. But in formula (25) for  $P_s(\vec{\lambda}_s)$ , which is associated with (29), a dependence of the information accumulation rate, with regard to the vector  $\vec{\lambda}_s$ , from the strategies  $\Gamma_i$ , does not exist, i.e., the information accumulation process is a passive or independent one.

By carrying out all the substitutions indicated above and then substituting (18) in (17), a final formula may be arrived at for the conditional, specific risk  $r_s$ . If the values of  $r_s$ , are considered in different experiments carried out with this system, then the vectors  $\vec{y}_s^*, \vec{u}_s$  and  $\vec{y}_{s-1}$ , which, generally speaking, are not known beforehand, may assume different values. Let  $P(\vec{y}_s^*, \vec{u}_{s-1}, \vec{y}_{s-1})$  be the density of the common distribution of these vectors; in such a case:

$$P(\vec{y}_s^*, \vec{u}_{s-1}, \vec{y}_{s-1}) = P(\vec{u}_{s-1}, \vec{y}_{s-1} | \vec{y}_s^*) P(\vec{y}_s^*) \quad (30)$$

In that case, the specific risk  $R_s$ , which represents the average value of  $r_s$ , where experiments have been conducted on a large scale, is determined by the formula:

$$R_s = M\{r_s\} = \int r_s P(\vec{y}_s^*, \vec{u}_{s-1}, \vec{y}_{s-1}) d\Omega(\vec{y}_s^*, \vec{u}_{s-1}, \vec{y}_{s-1}) \quad (31)$$

$$= \int r_s P(\vec{u}_{s-1}, \vec{y}_{s-1} | \vec{y}_s^*) P(\vec{y}_s^*) d\Omega(\vec{y}_s^*, \vec{u}_{s-1}, \vec{y}_{s-1})$$

Having substituted here the expression for  $r_s$ , the following formula is arrived at:

$$R_s = \int W_s[s, x_s^*(s, \vec{\lambda}_s), x_s] \cdot P_0(\vec{\lambda}_0) \cdot \prod_{i=1}^s P(\vec{\lambda}_i | \vec{\lambda}_{i-1}) \cdot P(x_s | \mu_s, u_s) \\ \cdot \prod_{i=0}^s P(y_i^* | \lambda_i) \cdot P_0(\vec{\mu}_0) \cdot \prod_{i=1}^s P(\vec{\mu}_i | \vec{\mu}_{i-1}) \cdot \prod_{i=0}^{s-1} P(y_i | \mu_i, i, u_i) \\ \cdot \prod_{i=0}^s \Gamma_i d\Omega(\vec{\lambda}_s, \vec{\mu}_s, x_s, \vec{y}_s^*, \vec{u}_s, \vec{y}_{s-1}) \quad (32)$$

It is important to note that, although in the given case, plant  $B$  has no memory, nevertheless risk  $R_s$ , in an  $s$  cycle, depends on all the  $\Gamma_i$  strategies at the moments  $t = 0, 1, \dots, s$ . The physical reason for this phenomenon, which is absent in an open loop system, is found precisely in the duality of control. Control at a  $k$  moment of time should be calculated not only with a view towards decreasing the specific risk  $R_k$ , which corresponds to this moment, but also towards promoting a risk reduction,  $R_i$  ( $i > k$ ), during the following moments, by means of a better study of the plant.



### Determination of the Optimum Strategy

In determining the optimum strategy<sup>8,9</sup> our thoughts are drawn towards dynamic programming<sup>1</sup>. Therefore introduce some auxiliary functions  $\alpha_k$  ( $0 \leq k \leq n$ ):

$$\begin{aligned} \alpha &= \alpha_k(\vec{y}_k^*, u_k, \vec{u}_{k-1}, \vec{y}_{k-1}) \\ &= \int_{\Omega(\vec{\lambda}_k, \vec{\mu}_k, x_k)} W_k[k, x_k^*(k, \vec{\lambda}_k), x_k] \cdot P_0(\vec{\lambda}_0) \cdot \prod_{i=1}^k P(\vec{\lambda}_i | \vec{\lambda}_{i-1}) P(x_k | \mu_k, u_k) \\ &\quad \cdot \prod_{i=0}^k P(y_i^* | \vec{\lambda}_i) \cdot P_0(\vec{\mu}_0) \cdot \prod_{i=1}^k P(\vec{\mu}_i | \vec{\mu}_{i-1}) \\ &\quad \cdot \prod_{i=0}^{k-1} P(y_i | \vec{\mu}_i, i, u_i) d\Omega(\vec{\lambda}_k, \vec{\mu}_k, x_k) \end{aligned} \quad (33)$$

Also, let:

$$\beta_k = \prod_{i=0}^k \Gamma_i \quad (34)$$

In that case, the formula for the risk  $R_n$ , which corresponds to the moment  $t = n$ , will assume the form:

$$\begin{aligned} R_n &= \int_{\Omega(\vec{y}_n^*, u_n, \vec{u}_{n-1}, \vec{y}_{n-1})} \alpha_n(\vec{y}_n^*, u_n, \vec{u}_{n-1}, \vec{y}_{n-1}) \beta_{n-1} \cdot \Gamma_n d\Omega(\vec{y}_n^*, u_n, \vec{u}_{n-1}, \vec{y}_{n-1}) \quad (35) \\ &= \int_{\Omega(\vec{y}_n^*, u_n, \vec{u}_{n-1}, \vec{y}_{n-1})} \beta_{n-1} \chi_n(\vec{y}_n^*, u_n, \vec{u}_{n-1}, \vec{y}_{n-1}) d\Omega(\vec{y}_n^*, u_n, \vec{u}_{n-1}, \vec{y}_{n-1}) \end{aligned}$$

where:

$$\begin{aligned} \chi_n(\vec{y}_n^*, u_n, \vec{u}_{n-1}, \vec{y}_{n-1}) \\ = \int_{\Omega(u_n)} \alpha_n(\vec{y}_n^*, u_n, \vec{u}_{n-1}, \vec{y}_{n-1}) \Gamma_n(\vec{y}_n^*, u_n, \vec{u}_{n-1}, \vec{y}_{n-1}) d\Omega(u_n) \end{aligned} \quad (36)$$

On the basis of the theorem of average value, taking (16) into consideration the following can be written:

$$\chi_n = (\alpha_n)_{av} \int_{\Omega(u_n)} \Gamma_n d\Omega(u_n) = (\alpha_n)_{av} \geq (\alpha_n)_{min} \quad (37)$$

Assume that all  $\Gamma_i$  ( $i < n$ ) are given and that the control process, right down to the moment  $t = n$ , has been realized. The selection of  $\Gamma_n$  must be in such a way as to minimize  $R_n$ . This may be accomplished if, for any arbitrary vectors  $\vec{y}_n^*$ ,  $\vec{u}_{n-1}$ ,  $\vec{y}_{n-1}$ ,  $\Gamma_n$  is selected in such a manner as to have function  $\chi_n$  minimal. Let  $\gamma_n$  equal  $\alpha_n$  and  $u_n^*$  be the value  $u_n$  that minimizes  $\alpha_n$ .

$$\gamma_n^* = \gamma_n(u_n^*) = \alpha_n(u_n^*) = \min_{u_n \in \Omega(u_n)} \alpha_n(u_n) \quad (38)$$

Evidently,  $u_n^*$  is the function of vectors  $\vec{y}_n^*$ ,  $\vec{u}_{n-1}$  and  $\vec{y}_{n-1}$ :

$$u_n^* = u_n^*(\vec{y}_n^*, \vec{u}_{n-1}, \vec{y}_{n-1}) \quad (39)$$

In that case, the optimum strategy  $\Gamma_n^*$  is given by the expression:

$$\Gamma_n^* = \delta(u_n - u_n^*) \quad (40)$$

where  $\delta$  is the unit impulse function. This denotes that  $\Gamma_n$  is the regular strategy, and not the random one, in which case the optimal value  $u_n = u_n^*$ . From (39) it is evident that the optimal value depends on the values previously observed by the control device  $A$ :  $u_s, y_s$  ( $s = 0, \dots, n-1$ ), and also on  $y_i^*$  ( $i = 0, \dots, n$ ).

It is very simple to prove the accuracy of expression (40). By substituting it in formula (35), one obtains, by virtue of the known property of the  $\delta$  function:

$$\chi_n = \min_{u_n \in \Omega(u_n)} \alpha_n(u_n, \vec{y}_n^*, \vec{u}_{n-1}, \vec{y}_{n-1}) = (\alpha_n)_{min} \quad (41)$$

But, according to (34), this is actually the lowest possible value for  $\chi_n$ . Consequently,  $\Gamma_n^*$  represents the optimum strategy.

In order to find the optimum strategies  $\Gamma_i^*$ , where  $i < n$ , one must shift gradually from the terminal moment  $t = n$  to the beginning—see references 8 and 9.

As a result, the rule stated below is arrived at for the determination of the optimum strategy  $\Gamma_n^*$ . Introduce the function:

$$\begin{aligned} \gamma_{n-k} &= \gamma_{n-k}(\vec{y}_{n-k}^*, \vec{u}_{n-k}, \vec{y}_{n-k-1}) = \alpha_{n-k} \\ &+ \int_{\Omega(y_{n-k}, y_{n-k+1}^*)} \gamma_{n-k+1}^*(u_{n-k+1}^*, \vec{y}_{n-k+1}^*, \vec{u}_{n-k}, \vec{y}_{n-k}) d\Omega(y_{n-k}, y_{n-k+1}^*) \end{aligned} \quad (42)$$

The magnitude  $\gamma_n^* = \alpha_n^*$ , according to (38). Now find the value that minimizes the function  $\gamma_{n-k}$ , in which case:

$$\gamma_{n-k}^* = \min_{u_{n-k} \in \Omega(u_{n-k})} \gamma_{n-k} = \gamma_{n-k}(u_{n-k}^*) \quad (43)$$

Evidently,

$$u_{n-k}^* = u_{n-k}^*(\vec{y}_{n-k}^*, \vec{u}_{n-k-1}, \vec{y}_{n-k-1}) \quad (44)$$

In that case, the optimum strategy is:

$$\Gamma_{n-k}^* = \delta(u_{n-k} - u_{n-k}^*) \quad (45)$$

i.e., the optimum strategy is regular and consists of the selection:  $u_{n-k} = u_{n-k}^*$ . From (44) it is evident that  $u_{n-k}^*$  depends on the values of  $u_i$  and  $y_i$ , which had been observed by the control device during the preceding moments, where  $i < n-k$ , and also on  $y_j^*$  ( $j \leq n-k$ ). Consequently, algorithm (44) is realized physically.

In the partial case, when the  $\vec{\lambda}_s$  process is converted to a random value of  $\vec{\lambda}$ , and  $\vec{\mu}_s$  to a random value of  $\vec{\mu}$ , formula (33) for  $\alpha_k$  is simplified and assumes the form:

$$\begin{aligned} \alpha_k &= \int_{\Omega(\vec{\lambda}, \vec{\mu}, x_k)} W_k[k, x_k^*(k, \vec{\lambda}), x_k] P_0(\vec{\lambda}) \prod_{i=0}^k P(y_i^* | \vec{\lambda}) \\ &\quad \cdot P_0(\vec{\mu}) \prod_{i=0}^{k-1} P(y_i | \vec{\mu}, i, u_i) P(x_k | \vec{\mu}, u_k) d\Omega(\vec{\lambda}, \vec{\mu}, x_k) \end{aligned} \quad (46)$$

If  $x_k^*$  are given in advance, then the formula proves to be still simpler:

$$\alpha_k = \int_{\Omega(\vec{\mu}, x_k)} W_k(k, x_k) P_0(\vec{\mu}) \prod_{i=0}^{k-1} P(y_i | \vec{\mu}, i, u_i) P(x_k | \vec{\mu}, u_k) d\Omega(\vec{\mu}, x_k) \quad (47)$$

These formulae had been previously brought out<sup>8,9</sup>.

### Examples

Consider three examples that illustrate the above theory. Figure 3 shows a representation of the simplest system for which  $h_s = 0$ ,  $\mu$  is the random quantity and the equations have the form:

$$\left. \begin{aligned} v_s &= u_s + g_s \\ y_s^* &= x_s^* + h_s^* \\ x_s &= v_s + \mu = u_s + g_s + \mu \end{aligned} \right\} \quad (48)$$

Let the stochastic quantities  $\mu$ ,  $g_s$  and  $h_s^*$  have normal distribution rules, with 0 average values and  $\sigma_\mu^2$ ,  $\sigma_{g_s}^2$  and  $\sigma_{h_s^*}^2$

variances, respectively. Further,  $\lambda_0$  and  $\sigma_\lambda$  are known quantities, in which case,

$$x_s^* = \lambda = \text{const}; P(\lambda) = \frac{1}{\sigma_\lambda \sqrt{2\pi}} \exp \left\{ -\frac{(\lambda - \lambda_0)^2}{2\sigma_\lambda^2} \right\} \quad (49)$$

Let:

$$W_s = W_s(s, x_s, x_s^*) = (x_s - x_s^*)^2 = (x_s - \lambda)^2 \quad (50)$$

As a result of basing the solution on the method described above (see reference 9) and with the application of formula (46), the optimal control rule is found in the following form:

$$u_s^* = \frac{\lambda_0}{1 + \left(\frac{\sigma_\lambda}{\sigma_g}\right)^2 (s+1)} + \frac{\sum_{i=0}^s y_i^*}{\left(\frac{\sigma_h}{\sigma_\lambda}\right)^2 + (s+1)} - \frac{\sum_{i=0}^{s-1} (x_i - u_i)}{s + \left(\frac{\sigma_g}{\sigma_\mu}\right)^2} \quad (51)$$

To explain the meaning of this formula, if interferences  $g_s$  and  $h_s^*$  were absent, then, for the purpose of obtaining an ideal value,  $x_s = x_s^* = \lambda$ , that would assure the magnitude  $W_s = 0$ , it would be necessary to establish the value of  $u_s = x_s^* - \mu = \lambda - \mu$ . In formula (51), the first two components yield an estimation for  $\lambda$  on the basis of the observed  $y_i^*$  values. The final term yields an estimation for  $\mu$  on the basis of the observed differences  $(x_i - u_i)$ . It is evident from Figure 3, in fact, that  $x_i - u_i = \mu + y_i$ . Consequently, an averaging of the differences  $(x_i - u_i)$  yields an estimation for the quantity  $\mu$ . With sufficiently high values for  $s$ , the final term of expression (51) is approximately equal to the arithmetic mean of the values  $(x_i - u_i)$ .

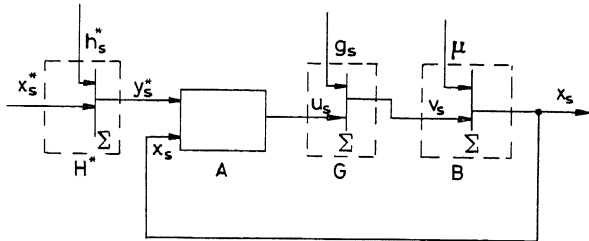


Figure 3

Consider another example pertaining to the same circuit in Figure 3. Let  $h_s^* = 0$ , while  $\mu$  is replaced by  $\mu_s$  and represents a Gaussian, discrete, Markov random process; in such case,

$$P_0(\mu_0) = \frac{1}{\sigma_0 \sqrt{2\pi}} \exp \left\{ -\frac{\mu_0^2}{2\sigma_0^2} \right\}$$

$$P(\mu_k | \mu_{k-1}) = \frac{1}{\sigma_1 \sqrt{2\pi}} \exp \left\{ -\frac{(\mu_k - \mu_{k-1})^2}{2\sigma_1^2} \right\} \quad (52)$$

The quantities  $g_s$  and  $W_s$  are the same as in the preceding example, and  $x_s^* = x^*$  is a known constant. In that case (see reference 10) the optimal control rule has the form:

$$u_k^* = x^* - \sum_{i=0}^{k-1} e_{i,k} (x_i - u_i) \quad (53)$$

The second component of formula (53) represents an estimation for the quantity  $\mu_s$ . The values of the weighting

coefficients,  $e_{i,k}$ , which are computed from comparatively complex formulae that are not set forth here, possess the property:

$$\frac{e_{i,k}}{e_{j,k}} < 1 \quad (0 \leq i < j \leq k) \quad (54)$$

The physical significance of this property in the optimum strategy (53) consists in that a lesser weight is imparted to information of older origin, inasmuch as it 'becomes obsolete'. Thus, there takes place, in the control device  $A$ , not only a process of storing new information, but also a process of degrading obsolete information.

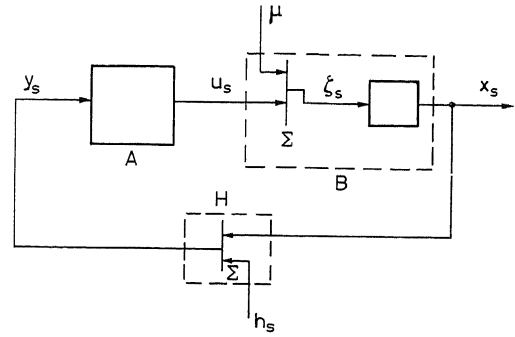


Figure 4

For the steady process, where  $k$  tends to infinity in formula (53), the  $e_{i,k}$  coefficients diminish in accordance with the law of geometric progression as the value  $v = k - i$  increases, i.e., as the previously measured difference  $(x_i - u_i)$  is withdrawn from the current moment of time. It is not difficult to realize such an algorithm by means of the simplest circuit.

The examples given above are degenerate, since they are equivalent to examples in which the value of the unknown parameter  $\mu$  is measured with a certain amount of error. The above theory, however, by means of a uniform method, makes it possible to examine even more complicated problems. Consider the system represented in Figure 4. The equations for this system have the form:

$$x_s = \zeta_s^2 = (u_s + \mu)^2, \quad y_s = x_s + h_s \quad (55)$$

The quantities  $g_s$  and  $h_s^* = 0$ . Noise  $h_s$  has a normal distribution with a zero average value and a variance  $\sigma_h^2$ . The quantity  $x_s^*$  is absent (for example, it may be assumed that  $x_s^* = 0$ ). All the values of  $W_s = 0$ , with the exception of the last one:

$$W_n = x_n \quad (56)$$

Assume that  $\mu$  is a random quantity with a probability density of  $P_0(\mu)$ , in which case  $|\mu| \leq 1$ . In the same manner, it may also be assumed that  $|u| \leq 1$ . The problem consists in determining the optimum algorithm for the control device  $A$  that would satisfy the condition:

$$R = M \{W_n\} = M \{x_n\} = \min \quad (57)$$

The solution of this problem produces the optimal method for searching a minimum of the parabolic function  $x = (u + \mu)^2$ , where  $\mu$  is unknown, and  $x$  is measured with an  $h$  error. At first, where  $i = 0, 1, \dots, n-1$ , probe values for  $u_i$  are established, and the corresponding quantities  $y_i$  are measured. Following this, where  $i = n$ , such a  $u_n$  is established as to give a minimal mathematical expectation to the  $x_n$  value that corresponds to it.

For the given problem:

$$P(y_i|\mu, i, u_i) = \frac{1}{\sigma_h \sqrt{2\pi}} \exp \left\{ -\frac{1}{2\sigma_h^2} [y_i - u_i^2 - 2u_i\mu - \mu^2]^2 \right\} \\ = \frac{1}{\sigma_h \sqrt{2\pi}} \exp \{ a_i + b_i\mu + c_i\mu^2 + d_i\mu^3 + \mu^4 \} \quad (58)$$

where:

$$\left. \begin{aligned} a_i &= -\frac{1}{2\sigma_h^2} (y_i - u_i^2)^2; & b_i &= -2u_i(u_i^2 - y_i) \frac{1}{\sigma_h^2} \\ c_i &= -\frac{1}{\sigma_h^2} (3u_i^2 - y_i); & d_i &= -\frac{2u_i}{\sigma_h^2} \end{aligned} \right\} \quad (59)$$

All  $\alpha_k = 0$ , with the exception of  $\alpha_n$ , for which, in conformity with (47) it is found that (integration for  $x_n$  is replaced by the substitution  $x_n = (u_n + \mu)^2$ ):

$$\alpha_r = \int_{-1}^1 (u_n + \mu)^2 P_0(\mu) \frac{1}{(\sigma_h \sqrt{2\pi})^n} \exp \left\{ \sum_{i=0}^{n-1} (a_i + b_i\mu + c_i\mu^2 + d_i\mu^3 + \mu^4) \right\} d\mu \quad (60)$$

By making use of this expression, it is possible to find  $\gamma_i$  and the values for  $u_i^*$ , which, in minimizing  $\gamma_i$ , prove to be optimal. This is accomplished by means of a succession of alternating minimizations and integrations, in the course of which, it is necessary to memorize the functions of three variables that are called, as is known (see, for example reference 11) sufficient coordinates, but the functions of three variables are too complex. For this reason, in the given case, the sufficient coordinates prove to be, figuratively speaking, insufficient for a convenient solution. However, the solution may be considerably simplified. As was indicated by calculations, by means of expansion into a Pike and Silverberg series<sup>12</sup>, it is possible to assume, with a sufficient degree of accuracy:

$$b_i\mu + c_i\mu^2 + d_i\mu^3 + \mu^4 = \\ \cong \varphi_1(y_i, u_i) f_1(\mu) + \varphi_2(y_i, u_i) f_2(\mu) \quad (61)$$

where  $\varphi_1, f_1, \varphi_2, f_2$  are some functions. In that case,

$$\alpha_n = \int_{-1}^1 (u_n + \mu)^2 P_0(\mu) \exp \left\{ A_{n-1} + \sum_{i=0}^{n-1} [\varphi_1(y_i, u_i) f_1(\mu) + \varphi_2(y_i, u_i) f_2(\mu)] \right\} d\mu \\ = \exp \{ A_{n-1} \} \int_{-1}^1 (u_n + \mu)^2 P_0(\mu) \exp \{ E_{n-1} f_1(\mu) + F_{n-1} f_2(\mu) \} d\mu \quad (62)$$

where

$$\left. \begin{aligned} A_s &= \sum_{i=0}^s a_i = A_{s-1} + a_s \\ E_s &= \sum_{i=0}^s \varphi_1(y_i, u_i) = E_{s-1} + \varphi_1(y_s, u_s) \\ F_s &= \sum_{i=0}^s \varphi_2(y_i, u_i) = F_{s-1} + \varphi_2(y_s, u_s) \end{aligned} \right\} \quad (63)$$

For this reason,

$$\gamma_n^* = \alpha_n^* = \min_{u_n \in \Omega(u_n)} \alpha_n = \exp \{ A_{n-1} \} \Theta_n^*(E_{n-1}, F_{n-1}) \quad (64)$$

where  $\Theta_n^*$  is the function of two variables.

Further,

$$\gamma_{n-1} = \int_{-\infty}^{\infty} \gamma_n^* dy_{n-1} = \exp \{ A_{n-2} \} \cdot \Theta_n(E_{n-2}, F_{n-2}, u_{n-1}) \quad (65)$$

and, if one assumes that:

$$\Theta_{n-1}^* = \min_{u_{n-1} \in \Omega(u_{n-1})} \Theta_{n-1} \quad (66)$$

$$\gamma_{n-1}^* = \exp \{ A_{n-2} \} \cdot \Theta_{n-1}^*(E_{n-2}, F_{n-2}) \quad (67) \\ (k=0, 1, \dots, n)$$

In a similar way, one obtains ( $k=0, 1, \dots, n$ ):

$$\gamma_{n-k}^* = \exp \{ A_{n-k-1} \} \cdot \Theta_{n-k}^*(E_{n-k}, F_{n-k}) \quad (68)$$

where

$$\Theta_{n-k}^*(E_{n-k}, F_{n-k}) = \min_{u_{n-k} \in \Omega(u_{n-k})} \Theta_{n-k}(E_{n-k}, F_{n-k}, u_{n-k}) \quad (69)$$

And so it is seen that it is only necessary to memorize the  $\Theta_{n-k}^*$  functions of two variables, which can be accomplished without any considerable difficulties. In minimization, it is sufficient to verify the extreme values of  $u_{n-k} = \pm 1$ .

## Conclusion

The dual control theory may be extended in various directions. Thus, for example, its extension to purely discrete systems, in which each of the quantities can assume only one of the permissible levels merits attention.

The development of this theory makes it possible to clarify the principles involved in the optimal teaching of discrete automatic machines.

The theory described above pertains to the 'Bayes' type, inasmuch as the assumption is made in it, that the *a priori* probability characteristics are known. However, the formulation of the dual control theory is likewise expedient for those cases where these characteristics are unknown. Such a formulation may be carried out either on the basis of the minimax principle or by the application of the idea of inductive probability.

At the present time, the most important problem for the immediate future is the development of approximate solution methods for dual control theory problems, the formulation of sub-optimal strategies, the determination of the numerical value of risk in quasi-optimal systems and its comparison with the value of risk in existing systems. Such a comparison will make it possible to clarify the extent of the gain that may be anticipated where we have a maximum degree of perfection in existing systems.

## References

- BELLMAN, R. *Dynamic Programming*. 1960. Inoizdat
- PONTRIAGIN, L. S., BOLTJANSKII, V. G., GAMKRELIDZE, R. V. and MISHCHENKO, E. F. *Mathematical Theory of Optimal Processes*. 1961. Fizmatgiz
- CHANG, S. S. L. *Optimum Synthesis of Control Systems*. 1961

- <sup>4</sup> FELDBAUM, A. A. *Calculating Devices in Automatic Systems*. 1959. Fizmatgiz
- <sup>5</sup> PUGACHEV, V. S. *Theory of Random Functions and its Application to Automatic Control Problems*. 1960. Fizmatgiz
- <sup>6</sup> LANING, D. KH. and BATTIN, R. G. *Random Processes in Automatic Control Problems*. 1958. Inoizdat
- <sup>7</sup> MIDDLETON, D. *An Introduction to Statistical Communication Theory*. 1960
- <sup>8</sup> FELDBAUM, A. A. Teoriya dualnogo upravleniya. *Avtomatika i Telemekhanika*, No. 9, Pt. I (1960); No. 11, Pt. II (1960); No. 1, Pt. III (1961); No. 3, Pt. IV (1961)
- <sup>9</sup> FELDBAUM, A. A. O nakoplenii informatsii v zamknutnikh sistemakh avtomaticheskogo upravleniya. *Izvest. Akad. Nauk SSSR, OTN, Energetika i Avtomatika*, No. 4 (1961)
- <sup>10</sup> FELDBAUM, A. A. Ob optimaliom upravleniya markovskimi obektami. *Avtomatika i Telemekhanika*, No. 8 (1962)
- <sup>11</sup> STRATONOVIIYA, P. L. K teorii optimalnogo upravleniya. Dostatochniye koordinatui. *Avtomatika i Telemekhanika*, No. 7 (1962)
- <sup>12</sup> PIKE, E. W. and SILVERBERG, T. R. Designing mechanical computers. *Machine Design*, No. 7, Pt. I (1952), 131-137; No. 8, Pt. II (1952), 159-163

## DISCUSSION

L. F. KAZDA, *Electrical Engineering Dept., University of Michigan, Ann Arbor, Michigan, U.S.A.*

Professor Feldbaum is to be complimented on his fine paper, which presents an extension of the author's companion papers published in *Automat. Telemekh.* in 1960. I would like to ask the following questions:

- (1) Have any practical systems been analysed, utilizing the ideas presented in this paper, to evaluate the average risk in a specific operating situation?
- (2) Have you given any thought to the case when the plant contains time delays or memory?
- (3) Since for any reasonable size of plant one generally is limited by computer capacity, have you considered sub-optimal strategies or approximations to a physical plant to give a usable solution?

A. A. FELDBAUM, *in reply*

I am grateful to Professor Kazda for the interesting questions he presented in his discussion.

(1) We have not yet considered any practical systems but we have solved several theoretical problems which in principle are of considerable interest. The results proved to be simple and give us hope that in practice also the solution of these problems will not be too complicated.

(2) One of my colleagues developed the dual control theory as applied to plants containing time delay. This paper will appear in the near future in *Automat. Telemekh.* As for the plants containing memory, i.e. dynamic plants, they were investigated in some of my works.

(3) For several simple cases we have constructed optimal strategies, which proved to be extremely simple. They were obtained by the approximation of curves found on a digital computer for an exact optimal strategy.

S. PASZKOVSKI, *Institute of Automation, Polish Academy of Sciences, Warsaw, Poland*

Dual control theory presented in this paper is of great importance in controlling the plant with incomplete information on the plant's parameters or medium. I understand that the whole dual control problem can be subdivided into the three following problems:

- (1) The search of algorithm when the information is known (the search of decision strategy).

(2) Determination of methods for accumulating the information in unknown parameters.

(3) Determination of methods for selecting the control actions (taking into account losses and accumulation of information) in order to optimally approach the decision strategy.

In practical examples I obtained the following: the parts of the system which simulate (with accumulation) unknown parameters made it easier to realize dual control.

A general schematic of the system is shown in *Figure A*.

Let us assume that the automation operates according to the optimal algorithm in the case where complete information is supplied to the input. However, we do not have the complete information and, therefore, it is necessary to study unknown parameters of the medium and the plant. It is necessary to control the system in such a manner that the decision strategy could be obtained with minimal additional losses.

In actual examples, the plant has a finite number of states and the transient probabilities are the unknown parameters.

$$P\{x_{s+1}/x_s, u_s\}$$

An estimate of unknown parameters has been carried out over intervals. Such a method made it possible to realize dual control which proved to be close to the optimal.

A. A. FELDBAUM, *in reply*

A technically sound way to solve the problem of constructing a controller often consists, especially in complex cases, in dividing the overall control task into two problems:

- (a) Determination of the unknown parameters of the plant, which is performed by separate units, e.g., with the aid of an analogue.
- (b) Determination of the control action from the parameters found.

However, in the general case this method is not optimal. In the optimal solution of the problem the control has a dual nature and both the above problems are solved together. In the general case it is not possible to divide the optimal controller into two separate units of the kind mentioned above. It is, however, quite possible that such a division may, in many cases, make it possible to implement a strategy close to optimal; in some cases a strictly optimal strategy may even be possible.

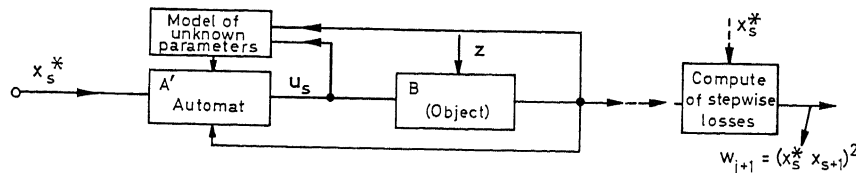


Figure A

H. CHESTNUT, *General Electric Company, 1 River Road, Schenectady, N.Y., U.S.A.*

To what extent does Professor Feldbaum find, for problems of practical systems, that analytical methods of solution are applicable, and to what extent are machine methods (computers) more satisfactory? Although not specifically referred to by Professor Feldbaum in this paper, there is the question to what extent the system designer permits the noise to be a significant factor and having to be taken into account, and to what extent is it worth while for him to try to reduce the noise to an acceptable amount.

A. A. FELDBAUM, *in reply*

In his comments Professor Chestnut mentioned two extremely interesting questions, on which the following can be stated:

- (1) The computers can be used in various ways:
  - (a) Obtaining the results of solution of a theoretical problem, for which a general solution specification is given; for example, in dual control theory problems of a non-trivial nature can be solved only with the aid of computers.
  - (b) Obtaining results by search in those cases when there are no analytical methods of solving the problem. For this one can use either universal digital computers or other equipment—for example, the automatic synthesis equipment designed at the Institute of Automation and Telemekhanics, Moscow.
  - (c) An optimal or sub-optimal algorithm, however found, can be put into a computer—digital, analogue, or hybrid—which in the given case will be the controller.

The proportion of purely analytical methods of problem solving is now comparatively small, and will evidently grow smaller. In the near future solving a problem will come to mean finding the programme by means of which the problem can be effectively solved on a computer.

(2) If the controlling plant itself, and also the meters and correcting devices, are specified—as is most often the case—the system designer is forced to take into account random interference and noise in the communications channels, as certain quantities with specified properties, which he cannot reduce at will.

Of course, he can, for example, put a filter in the feedback circuit leading from the plant output; however, filtering also causes an undesirable lag effect, and it is not clear in advance whether inclusion of a filter will lead to improvement or impairment of system behaviour.

Optimal choice of such a filter is also part of the synthesis of an optimal controller.

It is impossible to give general rules about when noise is significant and when it can be ignored. I can only give an example when even a small noise can have considerable influence. Optimal processes in a high-order system are only possible when the controller knows the state of the system, i. e. the values of the higher derivatives of its output. But how can they be obtained?

In many cases it is not possible to extract them from the controlled plant. All that can be done is to differentiate its output. However, in differentiation even the smallest noise mixed with the plant output signal leads to great noise in the values of the higher derivatives.

Therefore, it may turn out that failure to take noise into account in these cases will lead to a considerable deviation of processes from optimal.

A. M. HOPKIN, *University of California, Berkeley 4, California, U.S.A.*

Since application of this dual control calls for a rather elaborate controller, the practical problem of cost of the controller as compared to the advantage gained must be considered. I would like to ask Professor Feldbaum to what type of progress he foresees applying this technique, and what applications have already been made.

A. A. FELDBAUM, *in reply*

There are as yet no practical applications of the theory of dual control for the creation of real controllers.

The question of their cost is therefore still very much a matter of speculation. However, two things inspire hope that the cost will not be excessive:

(a) For some calculated cases sub-optimal strategies proved comparatively simple and realizable by means of uncomplicated circuitry.

(b) If the algorithm is realized by means of programme-control computers, making the algorithm more complex will only involve increasing computing time, not making the controller more costly. Therefore, if the computing time is acceptable, the cost of the controller does not depend on the complexity of the algorithm (until the complexity and the consequent slowdown of computations pass permissible limits, requiring a faster and more expensive computer).

B. FRIEDLAND, *Aerospace Research Center, 1225 McBride Avenue, Little Falls, U.S.A.*

(1) The distinction between the last two types (b) and (c) of optimal systems is not clear to me, particularly in view of the second paragraph in the second column of page 541.

(2) It appears that the extension of the approach of statistical decision theory represented by this paper to dynamic plants subsumes many of the problems now being considered in the theory of stochastic optimum control. Would the author care to discuss the following:

(a) Model used for dynamic processes; (b) Comparison of his results with those of other investigations in the U.S.S.R., U.K., and the U.S.A.?

A. A. FELDBAUM, *in reply*

(1) The difference between systems of types B and C is the same as that between observation and experiment. In B type systems it is only possible to observe the output of the source of random signals; it is not possible to act upon it. In C type systems, study of a plant with random properties can be forced by feeding to it control actions such that analysis of the reaction of the plant to these actions enables it to be studied faster.

Here the controller in essence performs 'experiments' with the plant, which cannot be done in an open-loop system.

(2) Extension of the theory of dual control to dynamic plants was made in another of my papers<sup>1</sup>. Comparison with the results of other approaches has been made only for certain very simple, so-called 'reducible' systems, i. e., systems for which an equivalent open-loop circuit can be built. The comparison showed the results to be identical.

As far as basically closed-loop systems are concerned, it is not possible to draw a comparison since such systems have not so far been studied.

Only one example has recently been examined<sup>2</sup>.

Comparison of the results of dual control theory with the results of studies in which two operations—determination of the parameters of the controlled plant and finding the control action from the parameter values obtained—are performed by separate units, is of undoubted interest.

## References

- <sup>1</sup> FELDBAUM, A. A. *Automat. Telemekh.* No. 1 (1961)
- <sup>2</sup> FLORENTIN, J. J. *Electronics and Control.* Aug. (1962)

D. XIROKOSTAS, *Electrical Engineering Department, University of Birmingham, Birmingham, England*

First, I would like to say that I found the paper very interesting and important because it gives a strict mathematical formulation and solution of the general discrete optimal control problem when nonlinearities, dynamics and random disturbances or noise are simul-

aneously present in the process; furthermore, because the presented theory, in my opinion, helps in the solution of some particular problems of this kind, it is of great practical importance. However, this is not indicated in the paper and I think more work in this direction is needed to clarify the field of application and the practical limitations of the theory.

The first point of my discussion refers to the last example, shown in Figure 4 of the paper, in which a non-linear process is considered, say of a simple parabolic form, and in which random disturbances  $\mu$  in the input and noise  $\xi_s$  in the output are simultaneously present. In the system, the optimum controller  $A$  derives informations about the process output  $x_s$  and uses it to produce the optimum control law in terms of the assumed performance criterion  $R$ . Although both random signals  $\mu$  and  $\xi_s$  have been assumed stationary, the effect of  $\mu$  sensed by the controller at the point  $H$ , because of the intermediate non-linearity, is a non-stationary signal, depending on the value of the control level  $u_s$ . Furthermore, this signal is not any more normal although this difficulty is not very important. Because of the first difficulty, however, the controller has to deal with a non-stationary signal depending on  $u_s$  which I think makes it difficult to derive the optimum control law for it. It is not clear to me how this has been taken into account in the paper, and I would like the author's comments on this point and ask if he has some more information about it, or has dealt with any particular example.

My second point is to ask if, in this particular example, process dynamics have been considered. In this case the problem becomes much more difficult, because the optimal decision at each step depends not only in the present state of the system, but also on previous ones, and so the Markoff process describing it is of higher order.

The solution of this last problem has been tried by the speaker, for the case that  $\mu$  disturbances are not present or they are simple step or ramp functions in an otherwise realistic model. The employed criterion is the same as the risk function  $R$  used in the paper. There have been two methods tried for this problem. One method is analytical, using the Markoff process theory. Because, in the problem, the Markoff process is of higher order, it is turned to that of first order, by introducing new variables and so extending the transition matrix of the process. Then the conventional theory of Markoff processes is applied.

The other method is a 'Monte Carlo' method. The whole system is simulated in a digital computer, making use of the facility of producing random numbers inside the computer of a desired distribution and variance. It is hoped that results from both methods will be published soon.

Thinking in terms of including the random disturbances  $\mu$  into the model, it seems that the above approach can be extended to the case where  $\mu$  consist of a Markoff process of first order and this will be the next step of the research.

It must be pointed out that the solution pursued above is sub-optimal because a stepping strategy with constant step in  $u_s$  is assumed. However, this assumption can be relaxed, keeping always the discrete nature of the controller, and so pursue optimal solution. Although this optimal solution is very interesting from a theoretical point of view and from the point of view of comparing its results with those of the previous solution, it does not seem that, at present, it can be easily applied to practical systems.

A. A. FELDBAUM, *in reply*

(1) Different idealizations of the system and formulations of the problem are possible. For example, in many problems of statistical theory, stationary random processes are examined. In the paper in question, however, another approach is used, and a stationary state is not required. In Example 3 the parameter  $\mu$  is a random quantity. The general formulae given in the paper enable one to find the solution specification if  $\mu$  is a Markov random process, not necessarily stationary.

(2) I fully agree that to take the dynamics of the plant into account makes the solution process complicated. In principle, however, the method of solution remains the same. A general solution specification, taking plant dynamics into account, was given by the author in *Automat. Telemekh.*, No. 1, 1961. I would be very interested to learn what Mr. Xirokostas has done to solve a similar problem.

O. L. R. JACOBS, *Department of Electrical Engineering, University of Edinburgh, Edinburgh, Scotland*

Professor Feldbaum has presented a very general method of formulating problems where it is necessary to learn about a process as it evolves in time. A rather academic, but classical example of this type of problem, is that known as the two-armed bandit.

I would like to ask Professor Feldbaum whether this problem has been formulated in the language of dual control theory and whether it has been possible to derive solutions to it.

A. A. FELDBAUM, *in reply*

Dr. Jacobs pointed out one interesting possible development of the theory of dual control. Bellman gave the principle of the solution of the two-armed bandit problem, using dynamic programming. An important feature of this problem is that the magnitude of the probability characterizing the plant is unknown, and only an *a priori* distribution is given for it.

For some classes of systems the theory of dual control has already been extended to similar problems, in which the *a priori* distribution of the unknown parameters of the plant contains unknown coefficients. In its turn the *a priori* distribution for these coefficients is specified.

The solution of the problem makes it possible to build an optimal adaptive or self-learning system, which accumulates information about these coefficients and clarifies their values.

P. DORATO, *Polytechnic Institute of Brooklyn, 333 Jay Street, Brooklyn 1, New York, U.S.A.*

I should like to ask Professor Feldbaum the following questions:

(1) Has any study been made of the case  $n = \infty$ , in particular has any study been made of the controllability of a system (in the sense that

$$\lim_{n \rightarrow \infty} \left[ \frac{1}{n+1} \sum_{s=0}^n R'_s \right]$$

exists) ?

(2) What ingredients must go into a control problem to make the strategy stochastic? Has an example of a stochastic strategy been worked out? If the strategy is stochastic how can it be implemented in a practical way?

A. A. FELDBAUM, *in reply*

(1) In the problem of optimal control of a Markov plant (Example 2 in the paper; see also *Automat. Telemekh.*, No. 8, 1962) an examination was also made of a steady-state process, in which the mean risk per cycle is reduced in the same sense as in the question.

(2) Generally speaking the optimal strategy will be random, if the optimum criterion is selected not in the form of the assembly average of the loss function, but in the form of its maximal value in the 'worst' case (a problem of the 'minimax' type). Nothing has so far been published in this direction. An optimal controller with random strategy can be implemented with the aid of a random signal generator and certain extra units. These devices are realized in the form of not very complex circuits.

# The Application of Random Test Signals in Process Optimization

P. M. E. M. van der GRINTEN

## Summary

In the determination and continuous adaptation of the optimum settings of a continuous process the influence of disturbances and dynamic effects can be ruled out by using random test signals superposed on the adjustable parameters.

In some cases use may be made of signals that are inherent in the process; in others it is necessary to superpose specially generated signals. The two methods are compared as regards possibilities of application, need of apparatus, and results.

## Sommaire

Dans la détermination et l'adaptation continue des régimes optimaux d'un processus continu l'influence des perturbations et des effets dynamiques peut être éliminée par l'utilisation de signaux d'essai aléatoires superposés aux paramètres réglables.

Dans certains cas, on peut utiliser des signaux inhérents au processus; dans d'autres il est nécessaire de superposer des signaux spécialement engendrés. Les deux méthodes sont comparées sous l'angle de leurs possibilités d'application, de l'appareillage nécessaire et des résultats.

## Zusammenfassung

Bei der Bestimmung der selbsttätigen kontinuierlichen Einstellung eines Optimums bei einem kontinuierlich arbeitenden Prozeß, lassen sich der Einfluß von Störungen und die dynamischen Auswirkungen ausgleichen, wenn man zufallsbestimmte Testsignale den einstellbaren Größen überlagert.

In einigen Fällen kann man hierfür die in dem Prozeß auftretenden Signale verwenden; in anderen Fällen ist es notwendig, eigens erzeugte Signale zu überlagern. Die Anwendungsmöglichkeiten, der technische Aufwand und die Ergebnisse der beiden Verfahren werden verglichen.

## Introduction

For the experimental maximization of a function

$$y = f(x_1, \dots, x_n) \quad (1)$$

it is advantageous to employ gradient methods, taking account of possible constraints. To be able to use such gradient methods the partial derivatives of  $y$  to  $x_1 \dots x_n$  should be known. If  $y$  be the profit function of a continuous process in which  $x_1 \dots x_n$  are the adjustable variables, these partial derivatives may be considered to represent the static gains of the process paths. For the experimental determination of these gains use may be made of various methods, all of which are based on the same principle, viz., making small changes in the  $x$  values and measuring the corresponding reactions of the  $y$  signal.

The main disadvantages experienced are (a) the mutual disturbance of the process paths and the influence of the inherent noise, and (b) the inclusion in the process of unknown and in some cases variable dynamic effects, which mask the pertinent static relation.

A solution to these difficulties may be found in the use of independent stochastic test signals for  $x_1 \dots x_n$ , which makes

possible the separate determination of the influence of each of the  $x$  variables on  $y$  separately by correlation. The dynamic effects can then be eliminated by using correlation functions<sup>1</sup> and not only correlation coefficients. These correlation functions do not contain any information about the phase of the signals, so that dead time and other non-minimum phase properties of the process do not complicate this measuring method, in contrast to methods using periodical test signals.

An elimination of dynamic effects as meant here is not pointless, because the profit is often a complicated function of a number of input and output variables, on account of which several parallel paths can develop in the process (= plant + profit calculation). This may give rise to non-minimum phase properties that are highly sensitive to small variations in process or settings. If no compensation is made for this phenomenon, it is no longer the pure gradient that is measured, so that the optimization may show a hunting effect. Consequently, complete elimination is particularly desirable whenever a moving optimum is to be followed.

## Random Test Signals

Starting from the existing relations between the auto-correlation and cross-correlation functions of various input and output variations, a general formula describing the static gain of the process paths was developed. For small variations the process is linear and correlation functions can indeed describe the dynamic properties of the process paths independently by means of convolution integrals. The relative deconvolution problem can be obviated if it is necessary to measure not the full dynamic properties but only the static gain, which can be found by extrapolation. This extrapolation is effected by integration of the correlation functions, as shown later.

A special case is given by

$$\frac{\Delta y}{\Delta x_n} \approx \frac{\int_{-\infty}^{+\infty} \phi_{x_n y}(\tau) d\tau}{\int_{-\infty}^{+\infty} \phi_{x_n x_n}(\tau) d\tau} \quad (2)$$

As  $\phi_{x_n x_n}(\tau)$  and  $\phi_{x_n y}(\tau)$  can be determined continuously, it is possible using this equation to measure the static gain of all process paths simultaneously. The expression has been successfully used for the continuous calculation of the gradient of the profit function of a gas burner, and a consecutive optimization<sup>1</sup>. The  $x$  signals were provided by the inherent spontaneous disturbances. In many cases, however, it will be necessary to use specially generated stochastic test signals, for the reasons given below.

(a) The spontaneous input variations may be interrelated, so that eqn (2) does not apply and the complete formula given at the end of the paper will have to be used.



(b) The spontaneous input variations may have an unsuitable power spectrum, which is shown by eqn (2) becoming indeterminate.

(c) The possibility of measuring the  $x$  signals may be lacking.

(d) Because of their wide amplitude and frequency ranges, the signals obtained from the spontaneous disturbances do not, generally, lend themselves well to being delayed for a sufficient length of time to enable the correlation functions to be calculated.

For these reasons a noise generator was designed, which had to generate a number of uncorrelated noise voltages having a stationary character and an adjustable power range. The output voltages are of a binary shape, so that they can be easily superposed on the setting parameters and can be delayed by means of punched-tape apparatus to enable the correlation to be determined. Some technical and theoretical details of this noise generator are given in the Appendix.

### Optimizers

The principal element of an optimizer operating according to the ideas outlined above will have to be a computer for eqn (2), i.e. for determining the integrals of correlation functions.

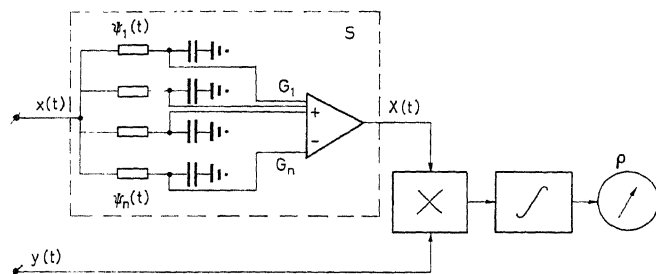


Figure 1. The continuous measurement of the integral of correlation functions

Consider the case in which use can be made of the inherent noise. The usual delay line for the calculation of correlation functions can be avoided by making use of a circuit as in Figure 1.

In this circuit the  $x$  signal is filtered by a number of RC filters with different time constants; special weights are applied to the output voltages which are then multiplied by  $y$ .

It can be demonstrated (see Reference 1 or the more simplified reasoning shown in the Appendix) that the result is actually an approximation of the integral of  $\phi_{xy}(\tau)$  for  $0 < \tau < T$ . The number of RC filters is determined by the accuracy required and by the magnitude of  $T$ .

Physically interpreted, the operation of the 'static filter'  $S$  consists in a determination of the moving average of the  $x$  signal. This may be expressed as

$$X(t) = \frac{1}{T} \int_0^T x(t-\tau) d\tau \quad (3)$$

The result,  $\rho$ , can now be calculated in a simple way:

$$\begin{aligned} \rho &= E^t [y(t) X(t)] = \frac{1}{T} \int_0^T E^t [x(t-\tau) y(t)] d\tau \\ &= \frac{1}{T} \int_0^T \phi_{xy}(\tau) d\tau \end{aligned}$$

Whether this expression, as it stands, can be used in eqn (2) will depend on the character of the correlation function before  $\tau = 0$  and after  $\tau = T$ . If before  $\tau = 0$  there are already significant values, these must be calculated separately with the help of an arrangement according to Figure 1, in which  $x$  and  $y$  have been interchanged. The 'averaging time'  $T$  must not be too short, as otherwise an appreciable part of the correlation function would be outside the area, so that the result would again be influenced by the dynamic properties. On the other hand,  $T$  should not be too long, as the signal to noise ratio decreases rapidly at higher  $T$  values, resulting in large measuring errors. It can be shown mathematically that eqn (2) becomes more nearly indeterminate as the  $T$  value increases.

Another method of performing the operation of eqn (3) on the stochastic input signal  $x$ , is the application of a sampler with period  $T$ , and a zero-order hold circuit. This can be shown by an analysis in the frequency domain. This suggestion makes it possible to determine the integrals of correlation functions, and hence the gradients of the response surfaces, by a simple manual calculation. The method described in the next section can be regarded as a mechanical and more efficient application of several of such samplers in parallel operation.

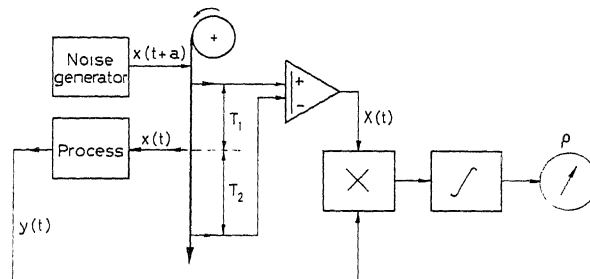
On the basis of the concept of the 'moving average' the case in which special stochastic signals are added is treated. With a circuit as shown in Figure 2 it is now possible to fix the limits of the moving average also before  $\tau = 0$ , by superposing a binary signal read from the pre-punched tape.

According to the diagram, the signal  $X(t)$  is given by

$$X(t) = \int_{-T_1}^{\infty} x(\tau-t) d\tau - \int_{T_2}^{\infty} x(\tau-t) d\tau = \int_{-T_1}^{T_2} x(\tau-t) d\tau \quad (5)$$

Correlation of this signal with  $y(t)$  again leads, according to eqn (4), to the desired integral.

Arrangements as shown in Figures 1 and 2 have been used in automatic optimizers: the first, for a process having strongly varying dynamic properties, with transients of the order of 1 min but without any dead time (a gas burner system); the second, for a process having transients of about 10 min and a dead time of 20 min, which rendered the RC filter method technically unsuitable. In both cases two input variables were optimized simultaneously by using the quantities calculated with the apparatus shown in Figures 1 and 2 respectively, for a control of these variables. Two methods for closing the optimizing loop were investigated. The first consists in a directly proportional control of the valves of the input variables by the output of pure integrators which follow upon the multipliers in Figures 1



(4) Figure 2. Use of a punched-tape apparatus for the determination of the integral of correlation functions



and 2. In the second method RC filters were applied instead of the pure integrators and the outputs were, after amplification, fed to limited-speed valve motors. In both methods the autocorrelation function of  $x(t)$  was assumed to be stable and was not taken into account. The first method proved to be slow but accurate, and is to be preferred for real optimizing problems. The second method is less accurate owing to inexact following

of the gradient, but fast and, hence, more useful in systems intended for the adaptive following of a moving optimum.

The results of some experiments performed with these optimizers according to the first method of closing the optimizing loop are shown in Figures 3 to 6. Figures 3 and 5 show a number of optimizing runs in a field of (estimated) lines of equal profit. Figures 4 and 6 show the relative time diagrams. The rate of

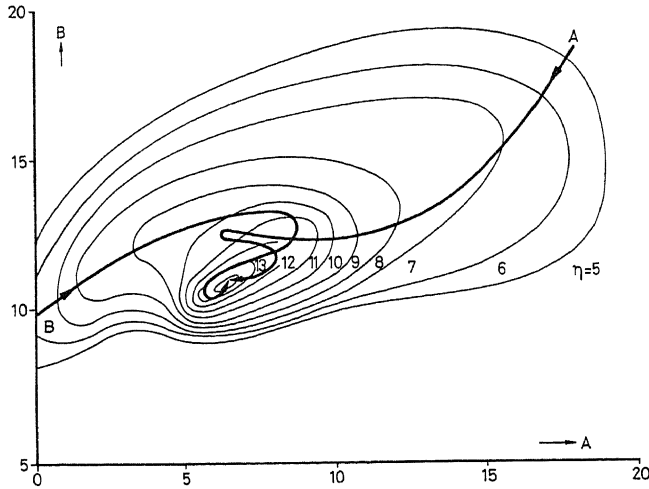


Figure 3. The optimization of a fast process by means of the method of Figure 1

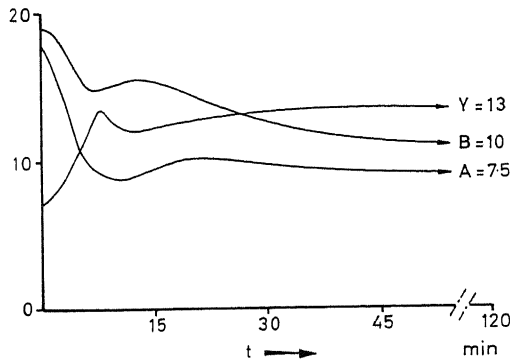


Figure 4. Time diagram of Figure 3

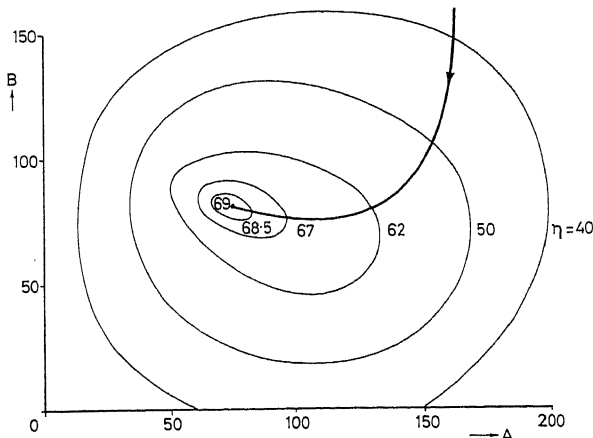


Figure 5. The optimization of a slow process by the method of Figure 2

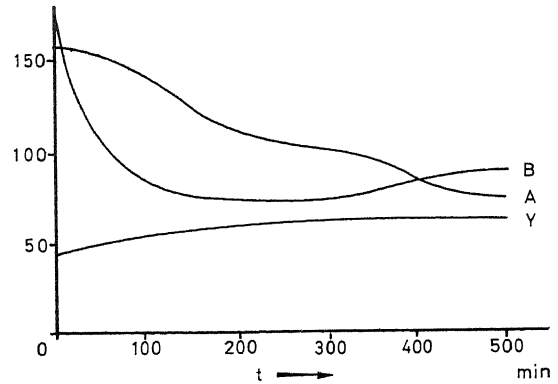


Figure 6. Time diagram of Figure 5

convergence is limited not only by stability requirements, e.g. the loop gain, but also by the accuracy of the measurement which is determined by the integration time. It can be seen here that with approximately equal loop gain of the optimizing circuits about the same convergence rates are obtained with respect to the transients of the processes, irrespective of the dead time and the exact character of the dynamic properties.

## Appendix

### The General Formula for the Static Gain of Process Paths

The relation between a random test signal on the  $n$ th adjustable input parameter and the  $y$  signal is given by the correlation function

$$\phi_{x_n y}(\tau) = \sum_{m=1}^N \int_0^{\infty} h_m(t) \phi_{x_n x_m}(\tau-t) dt \quad (6)$$

The test signal is here assumed to be so small that the relations of the process are linear.

If both members are integrated with respect to  $\tau$

$$\int_{-\infty}^{\infty} \phi_{x_n y}(\tau) d\tau = \sum_{m=1}^N \int_0^{\infty} h_m(t) \int_{-\infty}^{\infty} \phi_{x_n x_m}(\tau-t) d\tau dt \quad (7)$$

As  $h_n(t)$  is the response to a small pulse, the partial static gain (i.e. the final value of the response to a small step) can be calculated from eqn (7) by means of

$$\left( \frac{\Delta y}{\Delta x_n} \right)_{\text{stat}} = \int_0^{\infty} h_n(\tau) d\tau = \sum_{m=1}^N \alpha_{mn}^{-1} \int_{-\infty}^{\infty} \phi_{x_n y}(\tau) d\tau \quad (8)$$

in which the coefficient matrix  $\alpha_{mn}^{-1}$  is found as the inverse of the matrix

$$\alpha_{mn} = \left[ \int_{-\infty}^{+\infty} \phi_{x_n x_m}(\tau) d\tau \right] \quad (9)$$

In eqn (2) it was assumed that the input signals  $x$  are orthogonal, so that  $\alpha_{mn}$  is the unit matrix. Rudd<sup>2</sup> independently developed an analogous train of thought with respect to the frequency domain.

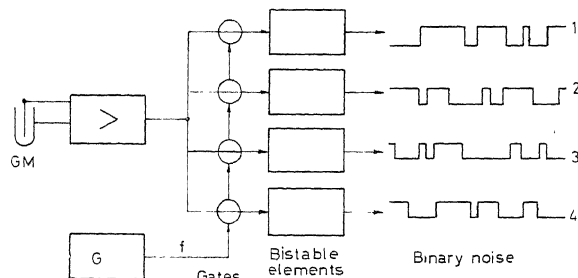


Figure 7. A generator for four statistically independent noise signals

#### The Design of Statistically Independent Noise Generators

The basis of the generator is a G.M. detector which, under the influence of a radioactive source, delivers pulses having a Poisson distribution (average  $\nu$  p.s.). These pulses pass through a number of parallel gates, which open  $f$  times per second in succession, for a short time  $T$ . The pulses passing through the gates also have a Poisson distribution\*, but their average number is much lower. The issuing pulses every time trip a bistable element. The correlation function of the block voltage thus obtained is rendered by

$$\phi_{xx}(\tau) = e^{-2\nu f T |\tau|} \quad (10)$$

The cross-correlation of signals from the channels is nil, as the pulses are completely independent of each other at their entry. Figure 8 gives measuring results for  $\phi_{x_1 x_1}$  and  $\phi_{x_1 x_2}$ .

#### Evaluation of the Coefficients $G_n$ in Figure 1

From Figure 1 it follows:

$$\rho = E^t [y(t) \cdot X(t)] = E^t [y(t) \cdot \int_0^\infty S(\tau) \cdot x(t-\tau) d\tau] \quad (11)$$

Here  $S(\tau)$  denotes the pulse response of the filter  $S$  in Figure 1:

$$S(\tau) = \sum_{n=1}^N G_n \frac{1}{R_n C_n} e^{-\tau/R_n C_n} \quad (12)$$

\* In actual fact this is not quite true: the pulses are always spaced by a multiple of  $T$ , but  $T \ll 1/\nu$ .

Eqn (11) is identical with eqn (4) if  $S(\tau) = 1/T$  in the range  $0 < \tau < T$  and  $S(\tau) = 0$  outside this range.

This shape of  $S(\tau)$  should now be approximated by a sophisticated choice of  $G_m$  in eqn (12). If eqn (12) is multiplied by  $1/R_m C_m \exp(-\tau/R_m C_m)$  and integrated with respect to  $\tau$

$$\int_0^\infty S(\tau) \frac{1}{R_m C_m} e^{-\tau/R_m C_m} d\tau = \sum_{n=1}^N G_n \int_0^\infty \frac{e^{-\tau(\frac{1}{R_n C_n} + \frac{1}{R_m C_m})}}{R_n R_m C_n C_m} d\tau \quad (13)$$

If the values of  $R_n C_n$  are fixed,  $G_m$  can be calculated from eqn (13). If, for example,  $N = 5$  and  $R_n C_n = 2^{-n}$ , the values of  $G_1 \dots G_5$  are 1.00, -1.34, 0.86, -0.34, 0.06.

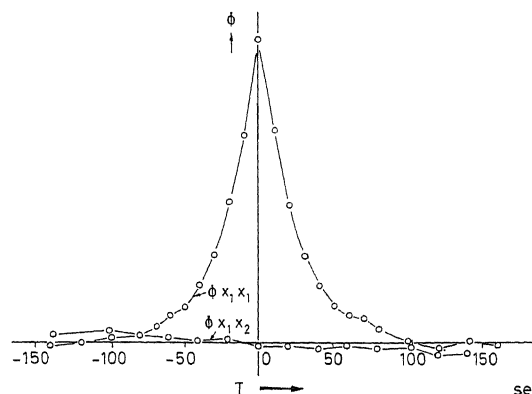


Figure 8. Correlation functions of the test signals generated by the method of Figure 7

#### Nomenclature

$y$	Profit function
$x$	Controllable variable
$\phi_{xy}(\tau)$	Correlation function of $x$ and $y$
$\tau, T, t$	Time (sec)
$h(t)$	Pulse response of a linear process
$\nu, f$	Frequency (c/sec)

#### References

- 1 VAN DER GRINTEN, P. M. E. M. *Congr. Eur. Fed. chem. Engrs* E 20 (1962) London; *Trans. Inst. chem. Engrs* 40, 6, 356
- 2 RUDD, D. F., ARIS, R., and AMUNDSON, N. R. *J. Amer. Inst. chem. Engrs* (1961) 376

#### DISCUSSION

R. PERRET, 44 Avenue Félix Viallet, Grenoble, France

May I ask the author to give further details about the last paragraph of his introduction: (a) About parallel paths that can develop in the process; (b) About the non-minimum phase properties that appear in these conditions.

P. M. E. M. VAN DER GRINTEN, in reply

By 'parallel paths in the profit calculation' I mean to say that profit is a function of both input variables and output variables. In such a case, a change in an input variable affects the profit along two different paths: directly and via the process. Take, for example, the gas burner system shown in Figure A.

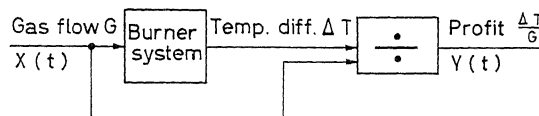


Figure A

The response of the profit to a stepwise increase of the gas flow is fast in relation to the  $G$ , but is slower in relation to the resulting temperature change  $\Delta T$ . There are now two possibilities [Figure B(i) and B(ii)].

In Figure B(i) there is a positive static gain, whereas in Figure B(ii) the gain is negative; apparently the two cases are on different sides

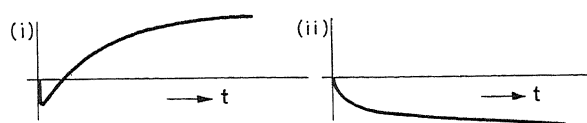


Figure B

of the optimum. The responses shown are typical of such non-minimum phase systems. The change of the apparent dynamics as a result of the settings of the input variables is also obvious from this example.

B. P. TH. VELTMAN, *Technological University of Delft, Lorentzweg 1, Delft, Netherlands*

Dr. van der Grinten apparently belongs to the most distinguished class of control mathematicians who have actually seen a process. The application of his method is very practical indeed—by asking the right question he eliminated the need for a digital computer making the methods also useful for small-scale plants.

My question concerns the type of binary test signals that are used. They are statistically independent. Orthogonal signals of the binary type could also be used; non-linearity in the process may, however, deteriorate the orthogonality. Is it possible to derive some conclusions about the non-linearity of the process from experiments with statistically independent and with orthogonal signals or are the relations too involved to be of any use? Could this comparison between both outcomes otherwise be recommended as a sensible means for establishing the linearity of the process?

P. M. E. M. VAN DER GRINTEN, *in reply*

I agree with your statement that slight non-linearities in the process will nullify the advantages offered by the use of orthogonal test signals. Therefore, I have stated that complete independence of the test signals is desirable. The use of orthogonal test signals for analysing the non-linearities is theoretically possible, but this would require knowledge of the dynamics.

H. A. BARKER, *The University, Glasgow W. 2, Scotland*

The author has presented a method for obtaining a linear expansion of a process output in terms of the process inputs. Has he considered

processes for which only a poor representation can be obtained in this way?

There is an alternative method for obtaining these coefficients for a linear expansion which is an optimum representation in the mean square error sense. If  $y$  is to be represented by the linear expansion

$$\sum_{i=1}^n a_i x_i$$

then the error  $e$  is given by

$$e = y - \sum_{i=1}^n a_i x_i$$

and is such that  $\overline{e^2}$  is minimum when the representation is optimum. Hence the coefficients  $a_i$  must be such that every average product function  $\overline{x_i e}$  is zero. The values of the coefficients may therefore be obtained by a method which is similar to that used by the author, but which is somewhat simpler by direct proportional control of the valve of each input variable  $x_i$  by the output of an integrator following a multiplier which has inputs  $x_i$  and  $e$ . This method has the advantage that it may be extended to give an optimum linear expansion for every process output in terms of linear or non-linear functionals of the process inputs.

#### References

- LUBBOCK, J. K. and BARKER, H. A. A solution of the identification problem. *Proc. JACC*, Minneapolis (June 1963)  
BARKER, H. A. The use of orthogonal functions for the solution of optimisation problems. *Ph. D. Thesis*, Cambridge University (1963)

P. M. E. M. VAN DER GRINTEN, *in reply*

The method of 'model updating' as proposed by Dr. Barker is certainly a realistic alternative for finding process parameters (see also the paper by Blandhol and Balchen). It has the advantage of storing the information and, in consequence, is capable of yielding faster results in certain cases.

On the other hand, the influence of process dynamics, when measuring the static gain, remains the main difficulty; the value  $\overline{X_i E}$  therefore gives no information on how to move the model-coefficients  $A_i$ . Maybe a combination of the method of my paper, that is the use of the moving averages of  $X_i$  and the method of the discussion, would be useful for this purpose.

# TECHNIQUES FOR SYSTEM STABILITY ASSESSMENT

## Eventual Stability\*

J. P. LASALLE and R. J. RATH

### Summary

In many real situations one wishes to consider the stability of states which are not equilibrium states. This immediately rules out Liapunov stability since his stability definitions imply that a stable state is an equilibrium state. A particularly good example is found in the study of stability of adaptive control systems. The desired state may not be an equilibrium state but may tend in time to act more and more like a stable equilibrium state. Such a stability, which here is called an eventual stability, is given a precise definition, and some basic properties of such stabilities are obtained. For example, if the system is autonomous or the state is an equilibrium state, then the eventual stabilities are the same as Liapunov stabilities, and the theorems can be viewed as generalizations of Liapunov's theorems. However, when the system is non-autonomous and the state is not an equilibrium state, then it is something new. Theorems are then given which show how Liapunov's direct method can be extended to study eventual stabilities and how qualitative estimates of the extent of the stability can be obtained. The authors believe this new concept of stability may play an important role in the theory and design of adaptive control systems and an example is given which illustrates how these ideas can be used in the design of an adaptive control system.

### Sommaire

Dans beaucoup de cas de la pratique on voudrait considérer la stabilité de certains états qui ne sont pas des états d'équilibre. Etant donné que les définitions de la stabilité selon Liapounov supposent que l'on a affaire à un état d'équilibre, on ne peut pas s'en servir dans ces cas. Exemple fort simple: étude de la stabilité des systèmes de commande adaptative. Quoique l'on n'ait pas affaire à un état d'équilibre, il se peut qu'avec le temps l'état considéré tende de plus en plus à agir comme ce dernier, le rapport décrit une telle stabilité comme «stabilité éventuelle» en donne une définition précise et en obtient certaines propriétés fondamentales. Ainsi si le système est autonome, ou bien si l'on a affaire à un état d'équilibre la définition de stabilité éventuelle coïncide avec celle de Liapounov, et les théorèmes correspondants peuvent être considérés comme des généralisations des siens mais hormis ces deux cas, il s'agit d'un fait nouveau. Le rapport donne alors des théorèmes montrant comment on peut étendre la méthode directe de Liapounov à l'étude de la stabilité éventuelle et

comment on arrive à en estimer qualitativement l'étendue. Les auteurs estiment que cette nouvelle stabilité est destinée à jouer un rôle important dans la théorie et la réalisation des projets de systèmes de commande adaptative les donnent un exemple montrant comment ces idées peuvent servir pour de tels projets.

### Zusammenfassung

In vielen konkreten Fällen möchte man die Stabilität von Zuständen, die keine Gleichgewichtszustände sind, betrachten. Dies schließt von vornherein die Anwendung der Liapunovschen Stabilitätsmethode aus, da diese Stabilitätsdefinitionen fordern, daß ein stabiler Zustand gleichzeitig ein Gleichgewichtszustand ist. Ein besonders gutes Beispiel liefert die Untersuchung eines selbststellenden Regelungssystems. Der gewünschte Zustand braucht kein Gleichgewichtszustand zu sein, aber er kann im Laufe der Zeit immer mehr wie ein Gleichgewichtszustand wirken. Für eine solche Stabilität — wir nennen sie „eventuelle Stabilität“ — wird eine genaue Definition gegeben, und es werden einige grundlegende Eigenschaften dieser Stabilitätsbegriffe gewonnen. Ist z. B. das System autonom oder der Zustand ein Gleichgewichtszustand, so sind unsere Begriffe der eventuellen Stabilität dieselben wie die von Liapunov, und die Sätze lassen sich als Verallgemeinerungen der Liapunovschen Sätze betrachten. Ist jedoch das System nichtautonom und der Zustand kein Gleichgewichtszustand, so liegen neue Gesichtspunkte vor. Hierfür werden Sätze angegeben, die auf eine Erweiterung der direkten Methode von Liapunov zur Behandlung der eventuellen Stabilität führen; sie gestatten eine quantitative Abschätzung des Stabilitätsgrades. Die Verfasser glauben, daß dieser neue Stabilitätsbegriff eine bedeutende Rolle für die Theorie und den Entwurf selbststellender Regelungssysteme spielen könnte; ein Beispiel zeigt, wie sich dieser Gedanke zum Entwurf eines selbststellenden Regelungssystems verwenden läßt.

### Introduction

The usual mathematical model for the stability problem in control is that the controlled system is described by a system of differential equations ( $\dot{x}_i = dx_i/dt$ )

$$\dot{x}_i = X_i(t, x_1, \dots, x_n, u_1, \dots, u_N), \quad i = 1, \dots, n$$

or in vector notation

$$\dot{x} = X(t, x, u)$$

(1)

\* This research was supported in part by the U.S. Air Force under Contract No. AF 49(638)-382, in part by the U.S. Army under Contract No. DA-36-034-ORD-3514 RD, and in part by the National Aeronautics and Space Administration under Contract No. NASr-103.

The general objective is to select the control  $u$  so that the system is maintained near some desired state. It is customary and convenient to take the desired state to be the origin  $x = 0$  (zero error). The control  $u$  will in general be a function of the state  $x$  of the system and of time, and with the selection of a specific control  $u$  eqn (1) becomes

$$\dot{x} = X(t, x, u(x, t)) = F(t, x)$$

To be a satisfactory control it is necessary that in some sense the system be maintained near the origin. In other words, the origin is in some sense to be stable.

It is accepted that stability or asymptotic stability means stability in the sense of Liapunov. There is today certainly no necessity for reviewing these definitions nor for describing Liapunov's second or direct method<sup>1-4</sup>. There is, however, one point that should be made. The definition of stability or asymptotic stability usually starts with the assumption that  $x = 0$  is an equilibrium state; that is,  $F(t, 0) \equiv 0$ . In fact, this is inherent in the definition of Liapunov stability: without the assumption that the origin is an equilibrium state it can be shown that the definition itself implies that the origin is an equilibrium state. Thus Liapunov stability applies only to states which are equilibrium states.

Now the kind of stabilities to be treated here are stabilities of states which are not necessarily equilibrium states but which as time passes tend to act more and more like stable or asymptotically stable equilibrium states. If the adaptive control problem is considered with some degree of realism, then this is the kind of stability that one would expect to achieve. If the plant or system being controlled is subjected to perturbations and a changing environment, then it is best to have the adaptive control make the desired state behave more and more like a stable equilibrium state or, better still, like an asymptotically stable state. Such stabilities are called 'eventual' stabilities and it is thought that they may play an important role in the theory and design of adaptive control systems.

In this paper concepts of eventual stability are defined, some of their basic properties are stated, and theorems which show how Liapunov's direct method can be extended to the study of eventual stability are given. If the system being studied is autonomous or if the origin is an equilibrium state, then these eventual stabilities are the same as Liapunov stabilities. In these cases the theorems given here are generalizations of Liapunov's theorems, but when the system is non-autonomous and the origin is not an equilibrium state, there is something new. In addition, theorems are given which show how qualitative results on the degree of eventual stability may be obtained; that is, an eventual region of asymptotic stability can be estimated. An example is given illustrating how the theory can be applied.

### Eventual Stability

The fundamental system is

$$\dot{x} = F(t, x) \quad (2)$$

and in some region  $\Omega$  it is assumed:  $t \geq 0$ ,  $\|x\| < R$ , uniqueness of solutions and a continuous dependence of the solutions on the initial condition. Here  $\|x\| = (x_1^2 + x_2^2 + \dots + x_n^2)^{1/2}$ , the Euclidean norm. When local properties are dealt with, it is always assumed that one is in  $\Omega$ . When properties in the large are treated,  $R = \infty$ . Let  $x(t, t_0, x^0)$  denote the solution of eqn (2) that starts at time  $t_0$  at  $x^0$ ; that is,  $x(t_0, t_0, x^0) = x^0$ .

### Definition 1

The origin is said to be eventually stable if, given  $\varepsilon > 0$ , there exist numbers  $\delta$  and  $T$  such that  $\|x^0\| < \delta$  implies  $\|x(t, t_0, x^0)\| < \varepsilon$  for all  $t \geq t_0 \geq T$ .

This is then the precise statement of what is meant by saying that 'as time goes on the origin tends to act more and more like a stable equilibrium state'. Eventual stability has also another intuitive meaning. If the origin is eventually stable, then the system has the property that if it has operated properly for a sufficiently long period of time it can be expected to continue to operate properly in the future. This interpretation of eventual stability is contained in the following result.

**Theorem 1**—Eventual stability of the origin is equivalent to the following: given  $\varepsilon > 0$  there exist numbers  $\delta$  and  $T$  such that  $\|x(t_1, t_0, x^0)\| < \delta$  for some  $t_1 \geq T$  implies  $\|x(t, t_0, x^0)\| < \varepsilon$  for all  $t \geq t_1$ .

Eventual stability is closely related to Liapunov stability and, in fact, the following is true.

**Theorem 2**—If the system of eqn (2) is autonomous ( $F(t, x) = f(x)$ ) or if the origin is an equilibrium state of eqn (2) ( $F(t, 0) \equiv 0$ ), then eventual stability of the origin is equivalent to uniform stability.

It is, however, easy to show that there are systems where the origin is eventually stable but not stable in the sense of Liapunov. A trivial example is  $\dot{x} = \phi(t)$ ,  $\phi(t) > 0$ ,  $\int_0^\infty \phi(t) dt < \infty$ .

A theorem which is similar to Liapunov's theorem on stability is now described. Let  $V(t, x)$  be a real-valued function with continuous first partial derivatives on  $\Omega$ . Then  $V(t, x)$  is said to be positive definite on  $\Omega$  if  $0 < u(\|x\|) \leq V(t, x)$  for all  $(t, x)$  in  $\Omega$ ,  $x \neq 0$ . Let  $\phi(t)$  be a function with the property that

$$\limsup_{t_0 \rightarrow \infty} \left\{ \sup_{a \geq 0} \int_{t_0}^{t_0+a} \phi(t) dt \right\} = 0 \quad (3)$$

This is, of course, a condition on the positive part of  $\phi$ . (For example, if  $\phi(t) \geq 0$  and if  $\int_0^\infty \phi(t) dt < \infty$ , then  $\phi(t)$  is a function of this type.) If  $V(t, x) \leq \phi(t)$  for all  $0 \leq \|x\| < R$ , then it will be said that  $V(t, x)$  is asymptotically non-positive in  $\Omega$ . Define

$$\begin{aligned} \dot{V}(t, x) &= \frac{\partial V}{\partial t} + (\text{grad}_x V) \cdot F(t, x) \\ &= \frac{\partial V}{\partial t} + \sum_{i=1}^n \frac{\partial V}{\partial x_i} F_i(t, x) \end{aligned}$$

It is then not difficult to show that:

**Theorem 3**—Let  $V(t, x)$  be positive definite in  $\Omega$  and have the property that  $V(t, x) \rightarrow 0$  as  $\|x\| \rightarrow 0$ , uniformly in  $t$  for  $t \geq 0$ . If  $\dot{V}(t, x)$  is asymptotically non-positive, then the origin of eqn (2) is eventually stable.

### Eventual Asymptotic Stability

As with eventual stability, eventual asymptotic stability is defined and an extension of Liapunov's asymptotic stability theorem for this type of stability is obtained.

### Definition 2

The origin of the system of eqn (2) will be said to be eventually asymptotically stable if (i) it is eventually stable and (ii) there is

an  $r > 0$  and a  $T_0$  such that  $\|x^0\| < r$  and  $t_0 \geq T_0$  imply  $x(t, t_0, x^0) \rightarrow 0$  as  $t \rightarrow \infty$ .

Condition (ii) states that if after a sufficiently long time the system is sufficiently close to the origin then it will tend to the origin as  $t$  approaches infinity. In analogy with Theorem 1 there is

**Theorem 4**—Condition (ii) of Definition 2 is equivalent to: there is an  $r > 0$  and a  $T_0$  such that  $\|x(t_1, t_0, x^0)\| < r$  for some  $t_1 \geq T_0$  implies  $x(t, t_0, x^0) \rightarrow 0$  as  $t \rightarrow \infty$ . Again as before:

**Theorem 5**—If the system of eqn (2) is autonomous or if the origin is an equilibrium state of eqn (2), then eventual asymptotic stability of the origin is equivalent to uniform asymptotic stability.

Turning now to the extension of Liapunov's theorem on asymptotic stability, it will be said that  $V(t, x)$  is asymptotically negative definite on  $\Omega$  if  $V(t, x) \leq \phi_r(t)$  for  $r < \|x\| < R$ , where for each  $r > 0$

$$\int_0^\infty \phi_r(t) dt = -\infty \quad (4)$$

**Theorem 6**—Assume that the origin is eventually stable and that  $V(t, x)$  is positive definite. If  $\dot{V}(t, x)$  is asymptotically negative definite, then the origin is eventually asymptotically stable.

In some respect this theorem is not as satisfactory as the theorem for Liapunov asymptotic stability. Here it is assumed to begin with that the origin is eventually stable and there is therefore the separate problem of establishing eventual stability. The reason for this is that what is called asymptotic negative definiteness does not, in general, imply asymptotic non-positiveness. A more useful result is the following:

**Theorem 7**—Assume that  $V(t, x)$  is positive definite on  $\Omega$  and that  $V(t, x) \rightarrow 0$  uniformly in  $t$  for  $t \geq 0$ . If  $\dot{V}(t, x) \leq -u(x) + h(t)$  on  $\Omega$ , where  $u(x)$  is positive definite on  $\Omega$  and

$$e^{-t} \int_0^t e^s h(s) ds \rightarrow 0 \text{ as } t \rightarrow \infty \quad (5)$$

then the origin of the system of eqn (2) is eventually asymptotically stable.

It was pointed out by Bojanic that eqn (5) is equivalent to

$$\limsup_{t \rightarrow \infty} \frac{1}{1+a} \left| \int_t^{t+a} h(s) ds \right| = 0 \quad (5a)$$

In the application of this theorem it is useful to note (see eqn (5a)) that condition (5) is satisfied if either.

$$h(t) \rightarrow 0 \text{ as } t \rightarrow \infty \quad (6)$$

or

$$\int_0^\infty h(t) dt = c \quad (c \text{ finite}) \quad (7)$$

As a consequence of the converse theorems<sup>5</sup> on the existence of Liapunov functions and Theorem 7, the following result is obtained almost immediately.

**Theorem 8**—Assume that the system

$$\dot{x} = X(t, x) \quad (8)$$

has a uniformly asymptotically stable origin. Then the system

$$\dot{x} = X(t, x) + p(t, x) \quad (8a)$$

is eventually asymptotically stable if  $|p(t, x)| \leq h(t)$  for  $\|x\| \leq r_0$  ( $r_0 > 0$ ), where  $h(t)$  satisfies eqn (5).

This theorem states that if the system of eqn (8a) approaches a uniformly asymptotically stable system sufficiently rapidly, then it is eventually asymptotically stable. Thus, for example, if  $\dot{x} = f(x)$  has an asymptotically stable origin and  $F(t, x) \rightarrow f(x)$  uniformly in  $x$  for  $\|x\| \leq r_0$  (some  $r_0 > 0$ ) as  $t \rightarrow \infty$ , then the origin of the system of eqn (2) is eventually asymptotically stable. Along these same lines there are also the following two theorems which give a method for obtaining qualitative information on the nature of the eventual asymptotic stability.

**Theorem 9**—Let  $M$  be a bounded closed set of  $n$  dimensional vectors  $x$  which contains the origin. Let  $N$  be a subset of  $M$  with the property that solutions of eqn (2) which start in  $N$  at a time  $t_0 \geq T$  remain thereafter in  $M$ . Suppose that there can be constructed a real-valued function  $V(t, x)$  with the following properties:

- (a)  $V(t, x) \rightarrow v(x)$  as  $t \rightarrow \infty$  uniformly for  $x$  in  $M$ .
- (b)  $\dot{V}(t, x) \rightarrow -w(x)$  as  $t \rightarrow \infty$  uniformly for  $x$  in  $M$ .
- (c)  $v(x)$  and  $w(x)$  are positive definite for  $x$  in  $M$ .

Then there exists a  $T_0 > 0$  with the property that  $x(t, t_0, x^0) \rightarrow 0$  as  $t \rightarrow \infty$  for each  $x^0$  in  $N$  and each  $t_0 \geq T_0$ .

**Theorem 10**—Assume that the set  $M$  defined by  $v(x) \leq l$  is bounded and that (a), (b), and (c) of Theorem 9 are satisfied. For each  $\delta > 0$  let  $M_\delta$  be the set defined by  $v(x) \leq l - \delta$ . Then corresponding to each  $\delta > 0$  there is a  $T_\delta$  such that  $x^0$  in  $M_\delta$  and  $t_0 \geq T_\delta$  implies  $x(t, t_0, x^0) \rightarrow 0$  as  $t \rightarrow \infty$ .

Theorem 9 states that  $N$  is an approximate eventual stability region and its size gives a measure of the system's stability. Theorem 10, which is probably simpler to apply, actually states that  $M_\delta$  is an approximate stability region. As time goes on ( $\delta \rightarrow 0$ ,  $T_\delta \rightarrow \infty$ )  $M_\delta$  approaches  $M$  so that  $M$  itself can be considered to be an approximate eventual stability region.

## An Application

The example of adaptive control to be considered here is an extension of an idea due to Rang<sup>6</sup>. It is in some sense artificial but illustrates that if one has enough information (which might well be statistical) about the unknown functions in the system to be controlled (the plant) then adaptive control is possible. Use is made here of the following theorem, which is closely related to Theorem 10:

**Theorem 11**—For the system

$$\dot{x} = f(t, x, y) \quad (x \text{ an } r \text{ vector})$$

$$\dot{y} = g(t, x, y) \quad (y \text{ an } s \text{ vector})$$

there is a scalar function  $V(x, y)$  with the following properties:

- (1)  $V(x, y)$  is positive definite and has continuous first partial derivatives for all  $x$  and  $y$ .
- (2)  $V(x, y) \rightarrow \infty$  as  $\|x\|^2 + \|y\|^2 \rightarrow \infty$ .
- (3)  $\dot{V}(x, y) \leq -W(x) + h(t)q(x, y)$  where  $W(x)$  is positive definite and continuous for all  $x$ ,  $q(x, y)$  is continuous for all  $x$  and  $y$ , and  $|h(t)|$  satisfies eqn (5) above.
- (4)  $x$  and  $y$  bounded imply  $f(t, x, y)$  is bounded for all  $t \geq 0$ .

Then the origin  $x = 0$ ,  $y = 0$  is eventually stable and corresponding to each  $r > 0$  there is a  $T_0$  such that  $\|x(t_0)\|^2 + \|y(t_0)\|^2 < r^2$  for  $t_0 \geq T_0$  implies that  $y(t)$  is bounded for  $t \geq t_0$  and  $x(t) \rightarrow 0$  as  $t \rightarrow \infty$ .

In the example to be considered in a moment  $x$  is the error and  $y$  is related to certain control parameters. What is wanted, which is the conclusion of the above theorem, is after the system has operated for a while that the error  $x$  go to zero and the control parameters remain bounded.

The general nature of the adaptive control system is shown in Figure 1. The differential equation of the plant is

$$\dot{x} = \alpha(t)Ax + \beta(t)p(t, x) + f$$

and the model is

$$\dot{y} = Ay + u(t)$$

The states  $x$  and  $y$  are  $n$  vectors,  $A$  is a constant  $n \times n$  matrix whose characteristic values all have negative real parts, and  $u(t)$  is continuous and bounded for all  $t \geq 0$ . For the plant it is assumed that  $p(t, x)$  is known and that it is bounded for all  $t \geq 0$  and all  $x$ . All that is assumed to be known about the scalar functions  $\alpha(t)$  and  $\beta(t)$  is that they are continuous and have bounded derivatives

$$\alpha(t) \rightarrow \alpha_0 \text{ and } \beta(t) \rightarrow \beta_0 \text{ as } t \rightarrow \infty$$

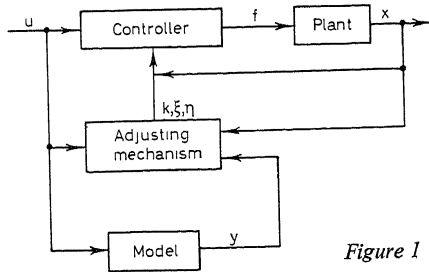


Figure 1

It is not assumed that  $\alpha_0$  and  $\beta_0$  are known. The problem is to determine the function  $f$  so that the plant follows the model ( $\varepsilon(t) = x(t) - y(t) \rightarrow 0$  as  $t \rightarrow \infty$ ) and the control parameters remain bounded. There are then good heuristic reasons for taking

$$f = k[(1 - \xi)Ax - \eta p(t, x) + u]$$

where  $\dot{k} = g_1(t, x, y, k, \xi, \eta)$ ;  $\dot{\xi} = g_2(t, x, y, k, \xi, \eta)$ ; and  $\dot{\eta} = g_3(t, x, y, k, \xi, \eta)$ .

It is desired, of course, to determine the functions  $g_1, g_2$ , and  $g_3$ . With  $\varepsilon = x - y$  the differential equation for  $\varepsilon$  is

$$\dot{\varepsilon} = A\varepsilon + (\alpha_0 - \xi)Ax + (\beta_0 - \eta)p + (k - 1)k^{-1}f + (\alpha - \alpha_0)Ax + (\beta - \beta_0)p$$

By assumptions on  $A$  it is known<sup>1</sup> that corresponding to each positive definite matrix  $C$  there is a positive definite matrix  $Q$  with the property that  $A'Q + QA = -2C$ . (Here and throughout the prime denotes transpose.) Select any positive definite matrix  $C$  and the corresponding  $Q$  and define

$$2V = \varepsilon'Q\varepsilon + c_1(k - 1)^2 + c_2(\xi - \alpha_0)^2 + c_3(\eta - \beta_0)^2$$

where  $c_1, c_2, c_3$  are positive constants. Then

$$\begin{aligned} \dot{V} = & -\varepsilon'Q\varepsilon + (k - 1)(c_1g_1 + k^{-1}f'Q\varepsilon) \\ & + (\xi - \alpha_0)(c_2g_2 + \varepsilon'Q\varepsilon - y'A'Q\varepsilon) \\ & + (\eta - \beta_0)(c_3g_3 - p'Q\varepsilon) \\ & + (\alpha - \alpha_0)(y'A'Q\varepsilon - \varepsilon'Q\varepsilon) \\ & + (\beta - \beta_0)p'Q\varepsilon \end{aligned}$$

Selecting

$$\begin{aligned} g_1 = & -c_1^{-1}k^{-1}f'Q\varepsilon \\ g_2 = & -c_2^{-1}(\varepsilon'Q\varepsilon - y'A'Q\varepsilon) \\ g_3 = & c_3^{-1}p'Q\varepsilon \end{aligned}$$

the following is obtained

$$\dot{V} = -\varepsilon'Q\varepsilon + (\alpha - \alpha_0)(y'A'Q\varepsilon - \varepsilon'Q\varepsilon) + (\beta - \beta_0)p'Q\varepsilon$$

It is not difficult to see that all the conditions of Theorem 10 are satisfied. The conclusion then is that  $x = 0, k = 1, \xi = \alpha_0, \eta = \beta_0$  is eventually stable and given  $r > 0$  there is a  $T_0 > 0$  such that

$$\|\varepsilon(t_0)\|^2 + (k(t_0) - 1)^2 + (\xi(t_0) - \alpha_0)^2 + (\eta(t_0) - \beta_0)^2 < r^2$$

for any  $t_0 \geq T_0$  implies that  $\varepsilon(t) \rightarrow 0$  as  $t \rightarrow \infty$  and  $k(t), \xi(t), \eta(t)$  are bounded for  $t \geq t_0$ . This, although not as conclusive an answer as might be desired, does say that in time the region in which satisfactory performance is obtained becomes the whole space. Without more information about the unknown functions which describe the plant this is perhaps the best that can be done. If  $\alpha(t) \equiv \alpha_0$  and  $\beta(t) \equiv \beta_0$ , then all solutions  $\varepsilon(t), k(t), \xi(t), \eta(t)$  are bounded and each  $\varepsilon(t) \rightarrow 0$  as  $t \rightarrow \infty$ .

## References

1. LASALLE, J. P., and LEFSCHETZ, S. *Stability by Liapunov's Direct Method with Applications*. 1961. New York; Academic Press
2. HAHN, W. *Theorie und Anwendung der direkten Methode von Liapunov, Ergebnisse der Mathematik und ihrer Grenzgebiete*. 1959. Berlin; Springer-Verlag (English translation in press)
3. KRASOVSKIĬ, N. N. *Some Problems in the Theory of Stability of Motion*. 1959. Moscow; Gosudarstv. Izdat. Fiz.-Mat. Lit. (English translation in press)
4. ZUBOV, I. V. *Methods of A. M. Liapunov and their Application*. 1959. Leningrad; Izdat. Leningrad University (English translation). 1961. Washington; AEC-tr-4439
5. MASSERA, J. L. Contributions to stability theory. *Ann. Math.* 64 (1956), 182
6. RANG, E. R. Adaptive controllers derived by stability considerations. 1962. *MPG Report 1529-TR 9*. Minneapolis; Honeywell

## DISCUSSION

R. TARJAN, *Hungarian Academy of Sciences, Budapest, Hungary*

The author's Definition 1 differs from the usual one ( $\|x'\| < \delta \Rightarrow \|x, t_0, x^0\| < \varepsilon$  for all  $t > t_0$ ) in that the condition for all  $t > t_0$  is replaced by the more general condition for all  $t_0 \geq t \geq T$ .

The crucial point seems to be the time interval 'T', after which the origin  $x^0$  tends to act as a stable state.

In practical cases the salient point seems to be to establish more or less explicit conditions for the existence and magnitude of  $T$ . This depends obviously on the parameters  $q$  of the system, which do not appear explicitly in eqn (2), the time dependence of which causes a drift of the origin  $x^0$ .

Further, an adaptive control system differs from a system with time-varying coefficients in that some of these coefficients are con-

trolled in a suitable manner in response to a suitable reference signal, which may be derived from either the input or the output, or a model or a combination of them. In this case, however, as has been shown previously<sup>1</sup>, we have two control loops within the system: the main loop working in a conventional manner, and the adaptive loop controlling some of the coefficients (parameters) of the main loop. The notion of the eventual stability refers obviously to the main loop, neglecting the adaptive loop. If, on the other hand, we are considering both loops simultaneously, a generalized equilibrium can be defined, which may or may not be stable according to the well-known stability varieties.

#### Reference

- <sup>1</sup> Proceedings of the Symposium on Self-Adaptive Systems, Rome 1962, in press

B. S. RAZUMICHIN, *Institute of Mechanics, Academy of Sciences, Moscow, U.S.S.R.*

The problem of the stability of motion is the problem of defining the conditions under which some solution of the system of differential equations that adequately describe the motion under investigation corresponds to the real existing motion. Thus not every solution of the system of differential equations describes the real motion (that is, the motion that can be realized). In this paper the definition of stability of some sets of the system states is described, but the set of states is not even a solution for the system of equations  $[F(t, 0) \neq 0]$ . For this reason I cannot agree that the stability definition given in the paper is a generalization of Liapunov's definition as the eventual stability is a development of Liapunov's definition for the class of fictional motions, which gives it a metaphysical character. The paper mentions some extremely important problems and contains proofs of theorems the correctness of which cannot be doubted. It seems to me, however, possible to set and solve the problems of the given class by the methods of the classical theory of Liapunov. For this purpose I can propose two methods.

(1) From the authors' definition it follows that the system of differential equations under investigation is such that there is one solution  $x_i(t)$  of the system, which approaches infinitely the zero point of coordinates, that is, such that for every  $\varepsilon > 0$  it is possible to determine a  $T > 0$  such that  $\|x(t)\| < \varepsilon$  with  $t > T$ . This is a real solution and the eventual stability is equivalent to the statement of the fact that there also exists the usual Liapunov's stability of such a solution.

(2) The system (1) considered in the paper

$$\dot{x}_i = F_i(t, x) \quad (1)$$

can be represented in the form

$$x_i = \phi_i(t, x) + F_i(t, 0) \quad (2)$$

where

$$\phi_i(t, 0) \equiv 0 \quad \text{and} \quad F_i(t, 0) \rightarrow 0 \quad (3)$$

The system (2) can be considered as a system with a constant acting disturbance which satisfies the conditions (3). The same problem was considered by I. G. Malkin and others in the limits of Liapunov's theory. Condition (3) permits proof of the asymptotic stability of system (1) under the assumption of uniform asymptotic stability of the trivial solution of the system  $\dot{x}_i = \phi_i(t, x)$  (this corresponds to Theorem 8 of the paper). In conclusion it may be said that the problem of the stability of the system infinitely approaching the equilibrium state can be solved in the limits of the classical Liapunov's theory and two ways are mentioned above to solve this problem. The first may

be recommended in cases when it is comparatively easy to find the solution which satisfies the condition  $x_i(t) \rightarrow 0$ . The second method can be recommended for all other cases.

J. P. LASALLE, *in reply*

The remarks that Dr. Razumichin has made are quite interesting although I must confess I am unable to understand his objections to considering eventual stability as a generalization of Liapunov stability. His inference that eventual stability is a 'development of Liapunov's definition for a class of fictional motions which gives it a metaphysical character' and his first proposal for overcoming his objections indicate that there is a misunderstanding.

Eventual stability of a state of a system is a simple and natural extension of Liapunov stability of a state. They both have to do with the asymptotic properties of solutions (motions) which start near the state. It is true that the motions are not the real physical motions of the system but the motions of the state of the system in state (phase) space. In this respect eventual stability is similar to Liapunov stability and, in fact, eventual stability includes as a special case Liapunov stability (Theorems 2 and 5). With the exception of some details of proofs and the generalization of conditions, the methods and concepts do not differ greatly from those of classical Liapunov theory.

In his second proposal, Dr. Razumichin brings out a point with greater clarity than we did in our paper. What he says is that if  $h(t) = |F(t, 0)|$  satisfies eqn (5) and if the origin of the system  $\dot{x} = F(t, x)$  is eventually asymptotically stable. This is the significance of Theorem 8: eventual asymptotic stability can be investigated by classical methods. Dr. Razumichin is also correct when he points out that Theorem 8 is closely related to total stability (stability under constantly acting disturbances).

It seems worth pointing out also that Theorem 11 is related to the important problem of the study of the stability of manifolds. It is a result of the eventual asymptotic stability of a noncompact manifold. Here the noncompact manifold is the  $r$ -dimensional linear subspace  $x = 0$ , which in applications to control systems corresponds to zero error. Considerations of how this may be applied suggest the need for a stronger and improved version of Theorem 11. With the same notation as in Theorem 11 a new result in this direction is the following.

Consider the system

$$\begin{aligned} \dot{x} &= f(t, x, y) & (x \text{ an } r \text{ vector}) \\ (*) \quad \dot{y} &= g(t, x, y) & (y \text{ an } s \text{ vector}) \end{aligned}$$

where  $f(t, x, y)$  is bounded for bounded  $x$  and  $y$  and all  $t \geq 0$ , and the following conditions on a Liapunov function  $V(x, y)$ :

(1)  $V(x, y)$  is positive definite and has continuous first partial derivatives

(2)  $V(x, y) \rightarrow \infty$  as  $\|x\|^2 + \|y\|^2 \rightarrow \infty$

(3)  $\dot{V}(x, y) \leq -W(x) + h_1(t)q(x, y) + h_2(t)V(x, y)$

where (i)  $W(x)$  is continuous and positive definite

(ii)  $q(x, y)$  is continuous

(iii)  $\int_0^\infty |h_i(t)| dt < \infty$ ,  $i = 1, 2$ .

**Theorem**—If for the system (\*) there exists a function  $V(x, y)$  satisfying (1)–(3), then the state  $x = 0$ ,  $y = 0$  is eventually stable and corresponding to each  $r > 0$  there is a  $T_r$  such that  $\|x(t_0)\|^2 + \|y(t_0)\|^2 < r^2$  for some  $t_0 \geq T_r$  implies  $y(t)$  is bounded and  $x(t) \rightarrow 0$  as  $t \rightarrow \infty$ .

If, in addition,

(4) For some  $K > 0$  and some  $0 < \alpha < 1$

$$|q(x, y)| \leq KV^\alpha(x, y)$$

then all solutions  $y(t)$  are bounded and all  $x(t) \rightarrow 0$  as  $t \rightarrow \infty$ .



# Non-linear Stability Analysis for Restricted Non-linearities Using the Second Method of Liapunov

H. NOUR ELDIN

## Summary

The methods treating the stability of a closed loop with single non-linearity lead to stability conditions that assure stability regardless of the shape of the non-linearity within a section in the non-linear plane. Thus, the stability criteria are always more than sufficient conditions for non-linearities that accept some restrictions on its shape. These restrictions must cause relaxation of the stability conditions. In this paper, this approach is used for both inherently stable and unstable control systems to get stability criteria that take into consideration the restrictions on the shape of the non-linearity.

## Sommaire

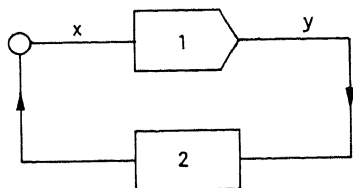
Les méthodes qui traitent la stabilité d'une boucle fermée contenant une seule non-linéarité conduisent à des conditions qui assurent la stabilité dans un secteur du plan non-linéaire sans considérer la forme de la non-linéarité. Ainsi, les critères de stabilité donnent toujours des conditions plus que suffisantes si les non-linéarités acceptent certaines restrictions pour leur forme. Ces restrictions permettent de relâcher les conditions de stabilité jusqu'à un certain point. Dans ce rapport ces considérations sont utilisées pour obtenir des critères pour la stabilité de systèmes dont la fonction de transfert est stable ou instable en prenant en considération des restrictions de la forme des non-linéarités.

## Zusammenfassung

Die Methoden, welche die Stabilität einer geschlossenen Schleife mit einer Nichtlinearität beschreiben, führen auf Stabilitätsbedingungen, die die Stabilität in einem Sektor der nicht-linearen Ebene garantieren, unabhängig von der Form der Nichtlinearität. Deswegen sind Stabilitätskriterien immer mehr als hinreichende Bedingungen, wenn die Nichtlinearitäten gewissen Beschränkungen in ihrer Form unterworfen sind. Diese Beschränkungen erlauben eine Milderung der Stabilitätsbedingungen. In obigem Artikel wird diese Betrachtungsweise angewendet, um für Regelungssysteme, deren offener Kreis stabil oder instabil ist, Stabilitätskriterien zu finden, welche die Beschränkungen in der Form der Nichtlinearität berücksichtigen.

## Introduction

The stability analysis presented in this paper deals with a single loop with single non-linearity. The block diagram of *Figure 1* shows the closed loop, for which the stability is studied in



*Figure 1. Block diagram of a single loop with single non-linearity without input function*

*Block 1. The non-linear element  $y = f(x)$  Block 2.  $G(s) = G_1(s) G_2(s)$*

absence of continuous external disturbing forces. The transfer function of the linear part (*Figure 1*) will be  $G(s)$ .

The linearized closed-loop system ( $y = kx$ ), may be one of the two types:

- (a) Inherently stable control system which is stable for any value of gain.
- (b) Inherently unstable, which is stable in a certain range of gain.

## Inherently Stable Control System

In this system, the open-loop poles and zeros of  $G(s)$  are in the left half of  $s$  plane. If the poles are simple, and their number is more than the number of zeros,  $G(s)$  can be expanded in partial fraction form.

$$G(s) = \frac{-X(s)}{Y(s)} = -\sum_{i=1}^{n+1} \frac{\alpha_i}{s + \lambda_i}$$

where  $Y = f(x)$ ,  $\lambda_i$  = pole of  $G(s)$  (real or in complex pairs, but different),  $\alpha_i$  = residues of  $G(s)$  at the corresponding pole, and  $n$  = order of the transfer function.

In this case, the transfer function  $G(s)$  can be replaced by  $n$  first-order transfer functions in parallel. The output variable after each first-order link  $z_i$  is related to  $f(x)$  by the relation

$$\frac{dz_i}{dt} = -\lambda_i z_i + f(x)$$

and to  $x$  by the relation

$$x = \sum_{i=1}^{n+1} \alpha_i z_i$$

$z_i$  is called the canonical variable.

The system equations can be written now in either one of the canonical forms

$$\frac{dz_i}{dt} = -\lambda_i z_i + f(x) \quad i = 1, 2, \dots, n$$

and

$$\dot{x} = \sum_{i=1}^n \beta_i z_i - r \cdot f(x) \quad (1a)$$

or

$$\frac{dz_i}{dt} = -\lambda_i z_i + f(x), \quad x = \sum_{i=1}^{n+1} \alpha_i z_i \quad (1b)$$

where

$$y = f(x) \\ \beta_i = -\alpha_i \lambda_i, \quad r = -\sum_{i=1}^n \alpha_i$$

One can get a simplified stability criterion using Luríé stability theorem. There are different simplified stability criteria<sup>1</sup>, in

which no conditions are imposed on the shape of  $y = f(x)$  except  $\dot{V}$  is negative definite if

$$f(0)=0, \quad \int_0^x f(x) dx \geq 0 \text{ or } xf(x) \geq 0 \text{ for all } |x| > 0$$

$$|a_{ij}| > 0 \quad \text{and} \quad \begin{vmatrix} A & -\theta \\ -\theta & 4rf_1 \end{vmatrix} > 0$$

$$z \sum_{j=1}^n \frac{a_{ij}}{\lambda_i + \lambda_j} + \beta_i f_1 = \theta_i \quad (7)$$

$$i = 1, 2, \dots, n$$

This means that the non-linearity is mainly in the first and third quadrant of  $[x, f(x)]$  plane, and passes through the origin.

There is no other restriction on the shape of  $f(x)$ . That is why the simplified stability criterion is more than a sufficient condition. One can relax the stability conditions by taking the shape of the non-linearity  $f(x)$  into consideration.

Consider as a Liapunov function

$$V = \sum_{i,j=1}^n \frac{a_{ij}}{\lambda_i + \lambda_j} z_i z_j + x^2/2 \quad a_{ij} = a_{ji} \quad (2)$$

which is positive definite everywhere for simple roots.

$$\dot{V} = Z'(z_i, x) B Z(z_i, x)$$

$$-B = \begin{vmatrix} & -\theta_1/2 \\ & a_{ij} \\ & \vdots \\ & -\theta_n/2 \\ -\theta_1/2 & \dots & -\theta_n/2 & rK \end{vmatrix}$$

$$\dot{V} \text{ is negative definite if } |a_{ij}| > 0 \text{ and } \begin{vmatrix} A & -\theta \\ -\theta & 4rk \end{vmatrix} > 0 \quad (3)$$

$$A = \|a_{ij}\|, \quad \theta = \theta_1, \theta_2, \dots, \theta_n$$

$$z \sum_{j=1}^n \frac{a_{ij}}{\lambda_i + \lambda_j} K + \beta_i = \theta_i \quad (4)$$

where

$$K = \frac{f(x)}{x} \quad i = 1, 2, \dots, n$$

In this criterion, the factor  $f(x)/x$  appears. One can find the range in which  $f(x)/x$  can satisfy the stability condition.

Again, consider the following Liapunov function

$$V = \sum_{i,j=1}^n \frac{a_{ij}}{\lambda_i + \lambda_j} z_i z_j + [f(x)]^2/2 \quad (5)$$

which is positive definite if  $f(0) = 0$ .  $f(x)$  may be in any of the four quadrants.

$$\dot{V} = Z'(z_i, f(x)) B Z(z_i, f(x))$$

$$-B = \begin{vmatrix} & -\theta_1/2 \\ & a_{ij} \\ & \vdots \\ & -\theta_n/2 \\ -\theta_1/2 & \dots & -\theta_n/2 & rf_1 \end{vmatrix} \quad f_1 = \frac{df(x)}{dx} \quad (6)$$

The general non-linearity  $f(x)$  must pass through the origin, and its slope has always to satisfy the above stability condition.

One can obtain a combination of the two criteria by adding the Liapunov functions. The new stability conditions will give restrictions on both  $K$  and  $f_1(x)$ .

The condition that  $f(x)$  must pass through the origin can be relaxed if we consider as Liapunov function

$$V = \sum_{i,j=1}^n \frac{a_{ij}}{\lambda_i + \lambda_j} z_i z_j + \frac{1}{2} [f(x) + \varepsilon]^2$$

which is positive definite if  $f(0) + \varepsilon = 0$ .  $\varepsilon$  is a constant.

The above stability conditions lead to the expansion of determinants of the order  $n+1$ . Similar stability conditions are derived which reduce the stability question to the expansion of  $n/2$  independent third order determinants.

Generally one can obtain, for a certain  $G(s)$ , the sufficient restrictions on the shape of the non-linear element that insure stability conditions.

If we use as Liapunov function

$$V = F(z_i) + x^2/2 + \frac{1}{2} \sum_{k=0}^n (f_k(x))^2 \quad (8)$$

where,

$$y = f_0(x)$$

the general non-linearity

$$f_K(x) = d^K f(x)/dx^K$$

and

$$F(z_i) = \sum_{i,j=1}^n \frac{a_{ij}}{\lambda_i + \lambda_j} z_i z_j \quad a_{ij} = \bar{a}_{ji}$$

or

$$= \sum_{i=1}^n \phi(a_i z_i, a_{i+1} z_{i+1})$$

like

$$= \sum_i \left[ \frac{a_i z_i \bar{z}_i}{\lambda_i + \bar{\lambda}_i} + \frac{a_{i+1} z_{i+1} \bar{z}_{i+1}}{\lambda_{i+1} + \bar{\lambda}_{i+1}} + \frac{a_{i,i+1} z_i \bar{z}_{i+1}}{\lambda_i + \bar{\lambda}_{i+1}} + \frac{\bar{a}_{i,i+1} \bar{z}_i z_{i+1}}{\bar{\lambda}_i + \lambda_{i+1}} \right]$$

$$i = 1, 3, \dots, n-1$$

In Appendix I, some stability conditions are shown for different restrictions on the shape of the general non-linearity  $f(x)$ .

### Inherently Unstable Control System

The study previously made for these systems introduces restriction on the amplification factor  $f(x)/x$ . Simplified stability criteria can be obtained either by pole or zero shifting<sup>1</sup>, or by using certain Liapunov functions that lead to stability determination if  $f(x)/x$  lies in a certain region. Generally, the linearized system is stable within the hatched area shown in Figure (2), and

only the cross-hatched area is stable for a general non-linearity having no restrictions except that it lies in this cross-hatched area. If more restrictions are imposed on the shape of  $f(x)$ , one can get stability criteria that extend the study of the system to the linearized stability region. The technique shown before can be applied for inherently unstable systems in conjunction with the pole-zero shifting method<sup>1</sup>, or by selection of certain Liapunov functions.

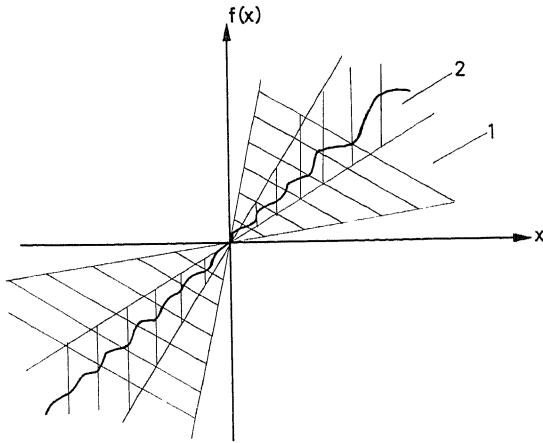


Figure 2. Region of stability for non-linear and linearized control system  
1. For linearized system 2. For non-linear operation

#### Pole and Zero Shifting Technique

The system equations are

$$y = f(x)$$

$$G(s) = -\frac{X(s)}{Y(s)} = -\frac{\sum_{i=0}^m a_i s^i}{\sum_{i=0}^n b_i s^i} \quad m < n$$

If  $G(s)$  has poles in the right-hand side of the  $s$  plane, one uses pole shifting by substitution

$$Y' = f(x) - c_p X = g(x) \dots \quad (9)$$

$$G'(s) = -\frac{X(s)}{Y'(s)} = -\frac{\sum_{i=0}^m a_i s^i}{\sum_{i=0}^n b_i s^i + c_p \sum_{i=0}^m a_i s^i} \quad (10)$$

If  $G(s)$  has zeros in the right-hand side of the  $s$  plane, we use zero shifting by substitution.

$$X' = X - c_z Y \quad (11)$$

$$G'(s) = -\frac{X'(s)}{Y(s)} = -\frac{\sum_{i=0}^m a_i s^i + c_z \sum_{i=0}^n b_i s^i}{\sum_{i=0}^n b_i s^i} \quad (12)$$

The values of  $c_p$  and  $c_z$  are chosen such that all poles and zeros are shifted to the negative half plane, which is the condition for linear stability.

For a pole shifted system, one can use the stability conditions in Appendix I. In this case the restrictions in these criteria will apply for  $g(x)$ . The constants  $\lambda, \alpha, \beta, r$  are those of the shifted transfer function  $G'(s)$  given by eqn (10).

In a zero shifted system, the number of poles in  $G'(s)$  are equal to the number of zeros. One has to use the modified first canonical form<sup>1</sup>

$$\frac{dz_i}{dt} = -\lambda_i z_i + f(x') \quad (13a)$$

$$x' = \sum_{i=1}^n \alpha_i z_i - c_z f(x') \quad (13b)$$

and

$$\dot{x}' = \sum_{i=1}^n \beta_i z_i - c_z \frac{df(x')}{dt} - r f(x') \quad (13c)$$

or

$$(1 + c_z f_1(x')) \cdot \dot{x}' = \sum_{i=1}^n \beta_i z_i - r f(x') \quad i = 1, 2, \dots, n$$

The values  $\lambda, \alpha, \beta, r$  are those of the original transfer function  $G(s)$ .

One can get stability conditions using the modified canonical form. However, this is not necessary as we can get the range of stability without using zero shifting.

#### Arbitrary Choice of Liapunov Function

The methods used before depend basically on the canonical form which requires the determination of the roots and residues. In the following, the general non-linearity is transferred directly in the system equations and one chooses the Liapunov function as if the non-linearity were already given in mathematical form.

#### Example 1

$$G(s) = \frac{1}{6} \frac{s+3}{s^2+2s-1} \quad (14)$$

This system was shown stable for general non-linearity<sup>1</sup> only if

$$|f(x)| \geq |6x| \quad \text{for all } |x| > 0, f(0) = 0$$

The system equations can be written in the form

$$x_1 = x$$

$$\dot{x}_1 = x_2$$

$$\dot{x}_2 = -(2 + f_1(x)/6)x_2 - (K/2 - 1)x_1 \quad (15)$$

let

$$V = \varepsilon x_1^2/2 + x_2^2/2 \quad \text{for } \varepsilon > 0 \quad (16)$$

$$dV/dt = (\varepsilon - K/2 + 1)x_1 x_2 - (2 + f_1(x)/6)x_2^2$$

Condition of stability will be

$$\left. \begin{aligned} 1 - K/2 + \varepsilon &= 0 & f(x)/x &> 2 \\ \text{and } 2 + f_1(x)/6 &> 0 & df(x)/dx &> -12 \end{aligned} \right\} \quad (17)$$

This shows that one can get a restricted non-linearity which covers all the stable region of the linearized system.

### Systematic Choice of Liapunov Function

Ingwerson<sup>2</sup>, has developed a powerful systematic attack to get a Liapunov function from the system differential equations. It is interesting here to get the conditions of stability of a second-order system using his systematic technique.

Let

$$G(s) = -\frac{X}{Y} = \frac{a_1 s + a_0}{s^2 + b_1 s + b_0} \quad y = f(x)$$

The system equation can be written in the form

$$\ddot{x} + (b_1 + a_1 f_1(x)) \dot{x} + (b_0 + a_0 K)x = 0 \quad (18)$$

Using the Ingwerson systematic technique to get  $V$  (Appendix III)

$$V = b_0 x_1^2 / 2 + a_0 \int_0^{x_1} f(x_1) dx_1 + x_2^2 / 2$$

which is positive definite if  $f(0) = 0$  and

$$b_0 x_1^2 / 2 + a_0 \int_0^{x_1} f_1(x_1) dx_1 \geq 0$$

therefore

$$b_0 + a_0 \frac{f(x)}{x} \geq 0$$

$$dV/dt = -(b_1 + a_1 f_1(x)) x_2^2$$

which is always negative if

$$b_1 + a_1 f_1(x) \geq 0$$

So, the conditions of stability of a non-linear second-order system with single non-linearity is that the coefficients of  $x$  and  $\dot{x}$  of equation (18) are always positive.

Applying this result to example (1), one gets the conditions

$$\frac{f(x)}{2x} - 1 \geq 0 \text{ and } 2 + f_1(x)/6 \geq 0$$

which are the conditions (17).

### Example (2)

$$G(s) = -\frac{x}{f(x)} = \frac{s-1}{(s+1)^2}$$

This system was solved using pole and zero shifting<sup>1</sup>, it is stable if

$$|0.5x| \leq |f(x)| \leq |x| \quad \text{for all } |x| > 0, f(0) = 0$$

Rewriting the differential equation we get

$$\ddot{x} + [2 + f_1(x)] \dot{x} + (1 - K)x = 0$$

The system is stable if

$$1 - f(x)/x \geq 0 \quad \text{i.e.} \quad |f(x)| \leq |x|$$

$$2 + df(x)/dx \geq 0 \quad \text{i.e.} \quad df(x)/dx \geq -2$$

This shows that there is no limit of  $f(x)$  so long as  $df(x)/dx \geq -2$ .

One notices that by these stability conditions for the second-order system one gets a stability region that covers all the stable

region of the linearized system. Using Ingwerson's method and writing the non-linearity with its derivatives in the system equations, one gets stability conditions for a higher order of  $G(s)$ .

### Stabilization of Non-linear Control

If one considers the second-order system in which

$$b_0 + a_0 f(x)/x < 0$$

one can get means for stabilizing it. Again considering the Liapunov function

$$V = b_0 x_1^2 / 2 + a_0 \int_0^{x_1} f(x_1) dx_1 + x_2^2 / 2$$

If relay characteristics are introduced for  $\dot{x}$  such that

$$x_2 = M \text{ sat}(\dot{x})$$

then the first condition for stability will be

$$\left. \begin{aligned} b_0 x_1^2 + a_0 \int_0^{x_1} f(x) dx_1 &\geq -M^2 / 2 \\ b_0 + a_0 (f(x)/x) &\geq -(M/x)^2 \end{aligned} \right\} \quad (19)$$

The system will be regionally stable in the region of  $x$  in which the above inequality applies. Thus, we have a stability condition that allows the system to operate in a region of the  $[f(x), x]$ -plane within which the linearized control is impossible.

### Conclusion

By imposing restrictions on the shape of a non-linearity in a single-loop control system, one can get sufficient stability conditions. The possible restrictions allowed to be defined determine generally the way in which the stability conditions are chosen. Two ways are described to get sufficient stability conditions and some of the conditioned stability criteria for inherently stable or unstable control systems are given.

### Appendix I

Conditioned stability criteria for systems described by the first canonical form of differential equations  $K = \frac{f(x)}{x}$  and  $f_1(x) = df(x)/dx$

$$\text{Re } \lambda_i > 0 \quad \theta = (\theta_1, \theta_2, \dots, \theta_n)$$

$$R = \|a_{ij}\| \quad a_{ij} = \bar{a}_{ij}, \quad i = 1, 2, \dots, n$$

No.

### Stability equation

$$1 \quad 2 a_i \sum_{j=1}^n \frac{a_{ij}}{\lambda_i + \bar{\lambda}_j} K + \beta_i = 0 \quad i = 1, 2, \dots, n$$

for all  $|x| > 0$

$$|a_{ij}| > 0 \quad \left| \begin{matrix} A & -\theta \\ -\theta & 4rk \end{matrix} \right| > 0$$

$$2 \quad \left| 2K \frac{a_i}{\lambda_i + \bar{\lambda}_i} + \beta_i \right| - 2 \sqrt{\frac{rka_i}{n}} < 0, \quad i = 1, 2, 3, \dots, n$$

for all  $|x| > 0 \quad a_i = 0$

$$3 \quad \left| 2 \frac{a_i}{\lambda_i + \bar{\lambda}_i} + f_1 \beta_i \right| - 2 \sqrt{\frac{r f_1 a_i}{n}} < 0 \quad i=1, 2, \dots, n$$

$$a_i > 0$$

$$4 \quad 2 \sum_{j=1}^n \frac{a_{ij}}{\lambda_i + \bar{\lambda}_j} + f_1(x) \beta_i = \theta_i$$

$$i=1, 2, 3, \dots, n$$

$$|a_{ij}| > 0, \quad \begin{vmatrix} A & -\theta \\ -\theta & 4 r f_1 \end{vmatrix} > 0$$

$$5 \quad 2 K \left[ \frac{a_i}{\lambda_i + \bar{\lambda}_i} + \frac{a_{i,i+1}}{\lambda_i + \bar{\lambda}_{i+1}} \right] + \beta_i = \theta_i,$$

$$2 K \left[ \frac{a_{i+1}}{\lambda_{i+1} + \bar{\lambda}_{i+1}} + \frac{a_{i,i+1}}{\lambda_i + \bar{\lambda}_{i+1}} \right] + \beta_{i+1} = \theta_{i+1}$$

for all  $|x| > 0$

$$i=1, 3, 5, \dots, n-1 \quad n \text{ even}$$

$$a_i > 0, \quad \begin{vmatrix} a_i & a_{i,i+1} \\ \bar{a}_{i,i+1} & a_{i+1} \end{vmatrix} > 0, \quad \begin{vmatrix} a_i & a_{i,i+1} & -\theta_i \\ \bar{a}_{i,i+1} & a_{i+1} & -\theta_{i+1} \\ -\bar{\theta}_i & -\bar{\theta}_{i+1} & \frac{8}{n} r k \end{vmatrix} > 0$$

$$6 \quad 2 \left[ \frac{a_i}{\lambda_i + \bar{\lambda}_i} + \frac{a_{i,i+1}}{\lambda_i + \bar{\lambda}_{i+1}} \right] + f_1 \beta_i = \theta_i,$$

$$2 \left[ \frac{a_{i+1}}{\lambda_{i+1} + \bar{\lambda}_{i+1}} + \frac{a_{i,i+1}}{\lambda_i + \bar{\lambda}_{i+1}} \right] + f_1 \beta_{i+1} = \theta_{i+1}$$

$$i=1, 2, 3, \dots, n-1 \quad n \text{ even}$$

$$a_i > 0, \quad \begin{vmatrix} a_i & a_{i,i+1} \\ \bar{a}_{i,i+1} & a_{i+1} \end{vmatrix} > 0, \quad \begin{vmatrix} a_i & a_{i,i+1} & -\theta_i \\ \bar{a}_{i,i+1} & a_{i+1} & -\theta_{i+1} \\ -\bar{\theta}_i & -\bar{\theta}_{i+1} & \frac{8}{n} r f_1 \end{vmatrix} > 0$$

## Appendix II

The conditions for  $-\beta$  to be positive definite are

$$|a_{ij}| > 0, \quad \begin{vmatrix} A & -\theta \\ -\theta & 4 r f_1 \end{vmatrix} > 0$$

The values of  $a_{ij}$  can be chosen to satisfy  $|a_{ij}| > 0$ . The second condition when expanded will take the form

$$D_1(a_{ij}) \cdot f_1^2 + D_2(a_{ij}) f_1 + D_3(a_{ij}) > 0$$

The above inequality will give a range of real  $f_1$  (for constant  $a_{ij}$ ) as long as  $D_2^2 - 4 D_1 D_3 > 0$ . By varying one of  $a_{ij}$  we get from the set the range in which the non-linear system is asymptotically stable.

If  $a_{ij} = 0$  for  $i \neq j$  the inequality reduces to

$$4 r f_1 > \sum_{i=1}^n \frac{|\theta_i|^2}{a_{ij}}$$

One gets a simpler inequality

$$\frac{4 r f_1}{n} > \frac{|\theta_i|^2}{a_{ij}}$$

which leads to  $\varphi$

$$\left| 2 \frac{a_{ij}}{\lambda_i + \bar{\lambda}_i} + f_1 \beta_i \right| - 2 \sqrt{\frac{r f_1 a_{ij}}{n}} < 0$$

The inequalities in Appendix I are applicable when  $x$  is a coordinate in the canonical form. If  $x$  is only an algebraic equation we use

$$\begin{vmatrix} \alpha_i \\ \vdots \\ a_{ik} & \alpha_k \\ \alpha_1 \dots \alpha_r 0 \end{vmatrix} < 0, \quad \begin{vmatrix} A & -\theta & \alpha \\ -\bar{\theta} & 4 r k & -1 \\ \alpha & -1 & 0 \end{vmatrix} < 0$$

$$k = \frac{f(x)}{x} \quad k=1, 2, \dots, n, \quad \alpha = (\alpha_1, \alpha_2, \dots, \alpha_r)$$

## Appendix III

$$\dot{x} = f(x)$$

$$x_1 = x$$

$$\dot{x}_1 = x_2$$

$$\dot{x}_2 = -[b_1 + a_1 f_1(x)] x_2 - (b_0 + a_0 K) x_1$$

$$\ddot{x} = B(x) \dot{x}$$

$$B = \begin{vmatrix} 0 & 1 \\ -[b_0 + a_0 f_1(x_1) + a_1 f_2(x_1) x_2] & -[b_1 + a_1 f_1(x_1)] \end{vmatrix}$$

$$\text{with } B'A + AB = -C$$

$$C = 2[b_1 + a_1 f_1(x_1)]$$

$$A = \begin{vmatrix} b_0 + a_0 f_1(x_1) + a_1 f_2(x_1) x_2 & 0 \\ 0 & 1 \end{vmatrix}$$

$$A(x_1, x_2) = \begin{vmatrix} b_0 + a_0 f_1(x_1) & 0 \\ 0 & 1 \end{vmatrix}$$

by integrating

$$\nabla V = b_0 x_1 + a_0 f(x_1) + x_2$$

integrating one gets

$$V = b_0 \frac{x_1^2}{2} + a_0 \int_0^{x_1} f(x_1) dx_1 + \frac{x_2^2}{2}$$

## Nomenclature

$x$	Input variable to the non-linear element
$X$	Input function to the non-linear element
$y$	Output variable from the non-linear element
$Y$	Output function from the non-linear element
$z$	The canonical variable
$f(x)$	The non-linear relation $y = f(x)$
$G(s)$	Linear transfer function
$s$	Complex variable

$\alpha$	Notation for residue
$\lambda$	Notation for pole
$i$	Index order
$j$	Index order
$a$	Arbitrary constant
$n$	Order of the system transfer function
$\varepsilon$	Constant
$K$	$f(x)/x$ the linearized gain
$f_K(x)$	$d^k f(x)/d x^k$

## References

- REKASIUS and GIBSON Stability analysis of non-linear control system by the second method of Liapunov. *I. R. E. Prof. Group AC.* (Jan. 1962)
- INGWERTSON A modified Liapunov method for non-linear stability analysis. *I. R. E. Prof. Group AC.* (May 1961)
- LETOV, A. *Stability in Non-linear Control Systems.* English edn 1961. Princeton; Princeton University Press
- HAHN *Theorie und Anwendung der Direkten Methode von Liapunov.* 1959. Springer Verlag

## DISCUSSION

G. P. SZEGÖ, *Research Institute of Advanced Studies, Baltimore, M.D., U.S.A.*

Between the stability equations obtained by the author, and those obtained by Lur'e, Letov and Rekasius and summarized by Lur'e<sup>1</sup>, there exists a major crucial difference. The stability equations obtained by the above-mentioned authors are indeed sufficient conditions for the absolute stability<sup>1</sup> of the system (1); as such, these conditions depend only upon the properties of the linear part of the system (1).

On the other hand, the stability equations presented in this paper contain explicitly the functions

$$K(x) = \frac{f(x)}{x} \quad \text{and} \quad f_1(x) = \frac{df}{dx}$$

and as such may be generally satisfied only by one particular non-linear system. However, in the case in which the non-linearity  $f(x)$  is known there exist methods<sup>2-5</sup> which allow a systematic construction of 'global Liapunov functions', i.e. Liapunov functions which describe the stability properties of any system in the whole state space. Even more severe shortcomings of this work are revealed by eqns (3) and (4), which can be satisfied only if  $K$  and  $f_1(x)$  are constant, i.e. for linear systems.

The problem of absolute stability of the system (1) with respect to certain restricted non-linearities is, however, very important. I shall solve it following the procedure of Kalman<sup>6</sup>.

Consider the control system  $\Sigma$

$$\begin{aligned} \dot{x} &= Ax - a\varphi(\sigma) \\ \sigma &= 2b'x \end{aligned} \quad (1)$$

and the continuous non-linear scalar function  $\varphi$ :

$$\varphi(0) = 0 \quad 0 < \varphi(\sigma) \leq \sigma^2 k \quad k < +\infty \quad (2)$$

In (1)  $A$  is an  $n \times n$  matrix,  $a$  and  $b$  are vectors. It is assumed that the matrix  $A$  is stable, that all linear systems in the class (1) are asymptotically stable and that the linear part of  $\Sigma$  is completely controllable and completely observable<sup>6</sup>.

We shall then represent  $\Sigma$  in its normal form<sup>5</sup>.

Consider the scalar function

$$v = x'Hx + \beta \int_0^\sigma \varphi(s) ds \quad h > 0 \quad (3)$$

whose total time derivative along the solutions of system (2) is

$$\begin{aligned} \dot{v} &= x'[A'H + HA]x - 2\varphi x'[Ha - \beta A'b - \alpha b] \\ &\quad - \varphi^2 \left[ \alpha \frac{1}{k} + 2\beta b'a \right] - \alpha \varphi \left[ \sigma - \frac{1}{k} \varphi \right] \end{aligned} \quad (4)$$

where the identity has been used

$$-\alpha [\varphi\sigma - 2\varphi x'b] = 0 \quad (5)$$

Let

$$\begin{aligned} A'H + HA &= -qq' \\ Ha - (\beta A' + \alpha)b + \gamma q \\ \alpha \frac{1}{k} + 2\beta b'a &= \gamma^2 > 0 \end{aligned} \quad (6)$$

If there exist a real scalar  $\gamma$ , a real vector  $q$  and a real matrix  $H = H'$  which satisfy the conditions (6), the scalar function (4) will take the form:

$$\dot{v} = -[x'q + \varphi\gamma]^2 - \alpha \left[ \sigma - \frac{1}{k} \varphi \right]^2 \quad (7)$$

By using the theorem proved by Kalman<sup>6</sup> one can show that for the existence of  $\gamma$ ,  $q$  and  $H$  satisfying eqn (6) it is necessary and sufficient that the conditions

$$\frac{\alpha}{k} + 2\beta b'a + 2 \operatorname{Re} \{ (\beta b'A + \alpha b')(Li\omega - A)^{-1}a \} \geq 0 \quad (8)$$

$$\frac{\alpha}{k} + 2\beta b'a > 0 \quad (9)$$

are satisfied for all real  $\omega$ .

The condition (8) may be rewritten in terms of the open-loop transfer function  $G(i\omega)$  of the linear part of  $\Sigma$  as follows

$$\frac{\alpha}{k} + \operatorname{Re} \{ (\beta i\omega + \alpha) G(i\omega) \} \geq 0 \quad (10)$$

**Theorem**—If (9) and (10) are satisfied for all real  $\omega$  and some real  $\alpha \geq 0$  and  $\beta$ , the system (1) is absolutely stable [globally asymptotically stable for all non-linearities of the class (2)].

**Proof.** From Kalman<sup>6</sup> it follows that for  $\alpha \geq 0$  the expression (7) is negative definite along the solutions of the system (1). To complete the proof one must show that  $v$  is positive definite. Since  $x'Hx > 0$  for  $\beta \geq 0$  this is trivially true. Let  $\beta < 0$ . Consider the minimum value of

$$\beta \int_0^\sigma \varphi(s) ds$$

in the class (2) which is for  $\varphi(\sigma) = k\sigma$ . It has been assumed that for  $\varphi(\sigma) = k\sigma$  the system (1) is asymptotically stable, hence  $v$  min. is positive definite. *Q.E.D.*

One can similarly operate in order to take into account the maximum derivative of the non-linearity  $\varphi = \varphi(\sigma)$  (Ezeilo<sup>7</sup>) into the stability conditions.

The case of indirect control may be treated in a similar way and the basic inequality (10) derived.

## References

- LUR'E, A. I. *Some Non-linear Problems in Theory of Automatic Control.* 1951. Moscow; Gostekhizdat
- ZUBOV, V. I. *The Methods of A. M. Liapunov and Their Application.* 1957. Leningrad; Isd-ro L.G.U. Leningrad

- <sup>3</sup> INGWERSON, D. R. *Trans. Inst. Radio Engrs. N.Y.* PGAC, 6 (1961), 199-201  
<sup>4</sup> SZEGÖ, G. P. *Trans. Amer. Soc. mech. Engrs. (D)*, 84 (1962), 571-578  
<sup>5</sup> SZEGÖ, G. P. *Trans. Amer. Soc. mech. Engrs. (D)*, 85 (1963), 137-142  
<sup>6</sup> KALMAN, R. E. *Proc. Nat. Acad. Sci.* 49 (1963), 201-205  
<sup>7</sup> EZEILO, J. O. C. *J. London Math. Soc.* 37 (1962), 405-409

J. C. GILLE, *Ecole Nationale Supérieure de l'Aéronautique, 32 Boulevard Victor, Paris, France*

The result indicated by the author on the stability of the second-order equation with non-linear dissipation term (in  $\dot{x}$ ) (18), may be generalized. Consider the more general non-linear equation of second order<sup>1</sup>

$$\ddot{x}f_1(\dot{x}) + \dot{x}f_2(x, \dot{x}, \ddot{x}) + xf_3(x) = 0 \quad (1)$$

and take the Liapunov function

$$V(x, \dot{x}) = \int_0^x f_1(\dot{x}) dx + \int_0^x f_3(x) dx \quad (2)$$

It is quite easily found that the derivative is

$$\dot{V}(x, \dot{x}) = -f_2(x, \dot{x}, \ddot{x})\dot{x}^2$$

which is never positive if

$$f_2(x, \dot{x}, \ddot{x}) > \varepsilon > 0$$

Further, the function  $V(x, \dot{x})$  being the sum of two integrals, is always positive if functions  $f_1(\dot{x})$  and  $f_3(x)$  are positive and if, moreover, they are single-valued, since an integral

$$\int_0^x f(x) dx$$

may be negative if  $f(x)$  is positive but not single-valued (Figure A). There is, in consequence, an unlimited stability for eqn (1) if (1)  $f_2$  is positive and (2)  $f_1$  and  $f_3$  are positive and single-valued.

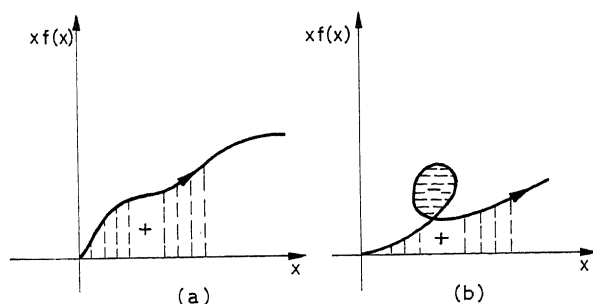


Figure A

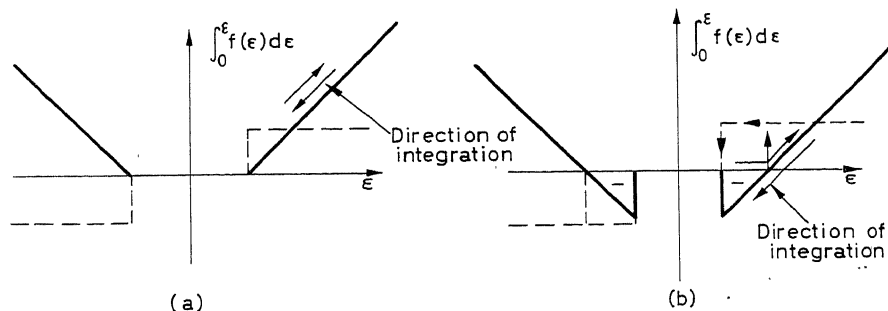


Figure B

The condition of single-valuedness of the characteristics of the non-linear element gives an explanation of the well-known phenomenon that a system composed of a linear element of second order and of a relay without hysteresis, is always stable, whilst the same system, but with the relay having hysteresis, may be unstable.

By virtue of the Liapunov function this is evident since the characteristic  $f(\varepsilon)$  of a relay without hysteresis is single-valued and, consequently, the integral

$$\int_0^\varepsilon f(\varepsilon) d\varepsilon$$

is always positive, whilst the characteristic  $f(\varepsilon)$  of a relay with hysteresis is not single-valued and the integral may be negative for certain values of  $\varepsilon$  as shown in Figure B.

Liapunov functions of type (2) may be generalized for certain equations of higher order. In this way, we may prove the stability of the equation of third order of the type

$$\ddot{\ddot{x}} + f_1(x, \dot{x}, \ddot{x})\ddot{x} + a\dot{x} + f_2(x, \dot{x}, \ddot{x})x = 0$$

if  $f_1$  and  $f_2$  are positive.

#### Reference

- <sup>1</sup> GILLE, J. C. and WĘGRZYN, S. Sufficient condition for the stability of non-linear systems of second order. *Automation* (November 1963)

M. AIZERMAN, *Institute of Automatics and Telemechanics, Moscow, U.S.S.R.*

I cannot understand this paper. The equations being solved [for example eqn (3)] serve to determine  $a_i$  coefficients of the quadratic form, but these equations have the term  $k = f(x)/x$ . For this reason their solutions do not give numbers but functions  $a_i(x)$ . However, in this case the Liapunov function (2) is not a quadratic form and it is not clear why the existence of real functions satisfying eqns (3) serves a definite, positive Liapunov's function (2). This remark leaves the rest of the paper not clear. In the best case we are in front of a hypothesis if eqns (3) have real solutions by any  $K_1 \leq K \leq K_2$  then an absolute stability exists.

Unfortunately this paper does not contain the proofs and offers no reason in favour of this hypothesis.

H. NOUR-ELDIN, *in reply*

I thank the discussors for their valuable remarks. I have added Appendix II to show that it is possible to obtain a range of asymptotic stability. I agree that by taking  $\theta = 0$ , ( $\theta_1 = \theta_2 = \dots = \theta_n = 0$ ) the inequality

$$\begin{vmatrix} A & \theta \\ \theta & 4rk \end{vmatrix} > 0$$

will lead to a system of equations and we cannot get a range of asymptotic stability in this case. I have already omitted such equations from the text.

# The Use of the Technique of Linear Bounds for Applying the Direct Method of Liapunov to a Class of Non-linear and Time-varying Systems

R. A. NESBIT

## Summary

By making use of certain inequalities, the stability of a class of systems is shown by the use of the basic theorems of the direct method. This set of inequalities defines a region, and if the non-linear or time-varying part of the system remains inside this region, then the stability of the system is assured. The usual methods of approximating the time response may be used, and thus this technique can be used to aid the design of control systems. The technique is applied to second- and third-order examples and is compared to the 'variable gradient' method of solving these problems.

## Sommaire

En utilisant certaines inégalités, ce rapport montre comment déterminer la stabilité d'une catégorie donnée de systèmes de réglage au moyen de la «méthode directe». L'ensemble de ces inégalités définit une certaine zone, et si la partie non-linéaire ou variant avec le temps se trouve à l'intérieur de cette zone, la stabilité du système est alors assurée. Les méthodes classiques d'approximation de la réponse en fonction du temps, peuvent être utilisées et complétées par cette nouvelle méthode, en vue de la détermination des systèmes de réglage. Cette méthode est appliquée à des exemples du 2e et du 3e ordre, et comparée aux résultats auxquels conduit l'application de la méthode des gradients variables aux problèmes de ce genre.

## Zusammenfassung

Durch Anwendung bestimmter Ungleichungen kann man die Stabilität einer Klasse von Systemen mittels der grundlegenden Theoreme der direkten Methode von Ljapunow prüfen. Diese Ungleichungen definieren einen Bereich; befindet sich der nichtlineare oder zeitveränderliche Teil des Systems innerhalb dieses Bereiches, so ist die Stabilität dieses Systems gesichert. Dabei kann man die üblichen Methoden zur Annäherung des Zeitverhaltens benutzen, wodurch sich dieses Verfahren für den Entwurf von Regelsystemen eignet. Dieses Verfahren wird auf Beispiele für Systeme zweiter und dritter Ordnung angewendet und mit der „variablen Gradienten-Methode“ zur Lösung dieser Probleme verglichen.

## Introduction

The direct approach to the study of the stability of motion which stems from the work of Liapunov<sup>2</sup> has been considered, presented and generalized in a large number of works, mainly by Russian and German investigators. Over 150 different references by 98 authors, published before 1958, are given by Hahn<sup>1</sup>. Of these works only about 11 were then available in English. At the end of 1959, Kalman and Bertram<sup>3, 4</sup> summarized the theory for English speaking readers.

Also, because of the growing interest in this general theory of non-linear systems many of the important Russian works are

now available in English<sup>5</sup> as well as many papers and some books written in English<sup>6</sup>. The principal works include those of Aizerman, Barbashin, Chetaev, Erugin, Hahn, Krasovski, Letov, Liapunov, Lurie, Malkin, Schultz and Zubov.

The direct method encompasses a number of theorems centred around so-called 'Liapunov functions' which are generalized 'distance' or 'energy' functions and around some precise definitions of stability. From a mathematical point of view the theory is both general and relatively complete. It is known that if a system is stable then there are many suitable Liapunov functions to prove stability, and the conditions on the Liapunov function for stability are concisely known. From the point of view of application of the theory to systems of interest to the control engineer, much has been done and much remains to be done. There are large classes of systems which are non-linear and/or time varying for which the stability behaviour can be determined by carrying out computations which have been prescribed already. Thus, it is worthwhile to study the canonical problems and their solutions (see the works of Lurie, Aizerman, and Letov).

However, until recently the application of the direct method has been impeded by the absence of a technique for constructing Liapunov functions. One of the first methods for constructing these functions is due to Zubov (see Hahn<sup>1</sup>) which requires solving a linear partial differential equation involving the gradient of the function. Recently Schultz<sup>7</sup> proposed a 'variable gradient method' of constructing Liapunov functions; this technique is based upon assuming the form of the gradient,  $\nabla V$ , and requiring single valuedness of  $V$ . The requirement that  $V$  be single valued (equivalent to  $\nabla \times \nabla V = 0$ ) allows the integration of the gradient along any convenient path to obtain the Liapunov function from the gradient. This construction method allows the application of the second method to many problems.

An alternative approach to assuming the form of the gradient of the function is to assume the form of the function itself or of its derivative. These choices are more obvious, but less fruitful. In the first case, integration of  $\dot{V}$  to find  $V$  is prohibitive, and in the other, it is difficult to make the assumed form of  $V$  general enough to give the required flexibility.

The stability theory can be applied in several ways. A particular system can be studied, or conditions on the system and its initial conditions for stability can be obtained. Not only can stability be determined, but also upper and lower bounds on the magnitude of the state vector can be obtained.

When the Liapunov function is known for one system, it may be used for other systems which are not too different. This fact has been used in many previous treatments; Hahn<sup>1</sup> uses linear bounds on the auxiliary term  $g(x)$  in the equation  $\dot{x} = Ax + g(x)$



with the Liapunov function for the linear equation  $\dot{x} = Ax$ . Antosiewicz<sup>8</sup> bounds the norm of the additional term  $g$  in the equation  $\dot{x} = f(x, t) + g(x, t)$  by the inequality  $\|g\| < h(t) \|x\|$  with  $h < 0$ ,  $\int_0^\infty h(t) dt < \infty$ . In the present paper linear bounds are used to assure the stability of the equation  $\dot{x} = f(x) + g(x, t)$ .

This work is strongly motivated by the similar treatment in Hahn<sup>1</sup> and it is felt that the slight generalizations are significant for the application of this theory to the design of control systems. Examples of the use of this theory are given for second and third-order systems with a linear principal term.

## Discussion

When a controller is designed for a system there is usually some uncertainty about the exact mathematical description of the forces which will act on the system. For example, in the design of an autopilot for an aircraft, there is an uncertainty of the exact functional relation between the angle of attack, the control surface deflection and the torques on the aircraft. The design is usually carried out by using the best available estimate of this relation and considering the linear model obtained from the first term of the Taylor's series expansion of the function. The design of the controller is then based upon this assumption and actually designed to operate for 'reasonable' changes from the model. It is desirable to have a controller which will control the system in a way which is insensitive to the actual torque relations that occur. This point of view leads to consideration of 'adaptive' controllers or the use of 'invariance' relations.

Using the theory of Liapunov it is possible to obtain estimates of the class of functions for which a given controller will operate satisfactorily. Estimates of this type are the subject of the problem of Aizerman, and the controllers of Lurie and Letov also seek classes of functions for which satisfactory operation is obtained.

The method for estimating the class of function in the problem of Lurie is concerned with one scalar function of one variable; the rest of the system is linear. In the following, bounds are obtained on a time-varying vector function of the state of the system. In the discussion of this method in Hahn<sup>1</sup>, bounds are obtained on the components of the non-linear function, and it is required to remain between fixed linear forms. That is,  $2n$  constant vectors are chosen such that either

$$x^T a_i \leq g_i(x) \leq x^T b_i \quad (i=2, \dots, n)$$

or

$$x^T a_i \geq g_i(x) \geq x^T b_i$$

with the vectors  $a_i$  and  $b_i$  chosen from a suitable set of vectors. This theory can be generalized somewhat to allow the use of a set of linear forms to bound the components. This generalization is important to the application of the theory because the function  $g(x)$  is thus not required to coincide with the intersection of the two hyperplanes. (The inequalities must be satisfied for all  $x$ . For those values of  $x$  for which  $x^T a_i = x^T b_i$  the requirement is quite stringent.) Also the components  $g_i$  can be functions of time as long as the linear bounds are satisfied for each instant.

## Mathematical Derivation

The dynamics of the system are represented in the  $n$  dimensional state space by

$$\begin{aligned} \dot{x} &= f(x) + g(x, t) & t \geq t_0 & & f(0) = 0 \\ x(t_0) &= x_0 & & & g(0, t) = 0 \end{aligned} \quad (1)$$

The 'norm' of the state vector  $\|x\|$  is given either by the Euclidian norm  $\|x\| = (x_1^2 + \dots + x_n^2)^{1/2} = (x^T x)^{1/2}$  or by some suitable norm such as  $\|x\| = (x^T B x)^{1/2}$  with a positive definite  $B$  matrix or  $\|x\| = (V(x))^{1/2}$  with  $V(x)$  positive definite. (This norm is not the usual norm of a linear space since the triangle inequality and the condition  $\|\alpha x\| = |\alpha| \|x\|$  are not necessarily satisfied.) Assume a Liapunov function  $V(x)$  is known for the system

$$\dot{y} = f(y) \quad (2)$$

with the property that its derivative for eqn (2),  $\dot{V}_2 = \nabla V \cdot f$ , is negative definite. Let  $X$  be a fixed open region containing the origin, let  $h$  be a fixed positive constant, and consider the set  $P_h$  of all  $n \times n$  constant matrices  $G$  for which

$$\nabla V \cdot f(x) + \nabla V \cdot Gx \leq -h \|x\|^2 \quad (3)$$

for all  $x \in X$ .

Let  $V_0$  be the largest region in state space which is bounded by  $V(x) = \text{constant}$  and contained in  $X$ .

The matrix  $G_\alpha$  can be considered to be composed of the vectors  $g_{i\alpha}^T$  ( $i = 1, 2, \dots, n$ ) with

$$g_{i\alpha}^T = (g_{i1}, g_{i2}, \dots, g_{in})_\alpha \quad (4)$$

$$G_\alpha^T = [g_{1\alpha}, g_{2\alpha}, \dots, g_{n\alpha}] \quad (5)$$

These vectors will be used to generate linear forms,  $x^T g_{i\alpha}$ .

A set of matrices  $R$  in  $G$  space will be called symmetric if the following condition is satisfied. If the  $i$ th row of one matrix in  $R$  ( $i = 1, \dots, n$ ) is replaced by the  $i$ th row of another matrix in  $R$  then the resulting matrix is also in  $R$ . A symmetric subset of  $P_h$  will be designated by  $R_h$ .

On the basis of these definitions the following theorem may be stated.

**Theorem of Linear Bounds**—If the following conditions are satisfied

(a)  $V(y)$  is a Liapunov function for eqn (2) with a negative definite derivative for  $y \in Y$ ;

(b) For each fixed  $(x, t)$  in  $(X, t \geq t_0)$   $2n$  vectors  $g_{i\alpha}$  and  $g_{i\beta}$  can be chosen from some  $G_\alpha$  and  $G_\beta$  in  $R_h$  such that the inequalities

$$x^T g_{i\alpha} \leq g_i(x, t) \leq x^T g_{i\beta} \quad (i=1, 2, \dots, n) \quad (6)$$

or

$$x^T g_{i\alpha} \geq g_i(x, t) \geq x^T g_{i\beta}$$

are satisfied, then the equilibrium position of eqn (1) has the same stability behaviour as eqn (2) in the small. In the case of asymptotic stability, the region  $V_0$  is included in the entrance region.

**Proof**—The inequalities of eqn (6) assure that  $V(x)$  is also a Liapunov function for the system of eqn (1). This can be shown by considering the derivative of  $V$  for eqn (1):

$$\dot{V}_1 = \nabla V \cdot f(x) + \nabla V \cdot g(x, t) \quad (7)$$

By multiplying the  $i$ th inequality from eqn (6) by the  $i$ th component of  $\nabla V$ , arranging the inequality with the greater term on the right, and adding the results for all  $i$ ,

$$\nabla V \cdot G_1 x \leq \nabla V \cdot g(x, t) \leq \nabla V \cdot G_2 x \quad (8)$$

where  $G_1$  and  $G_2$  are made up of rows from  $G_\alpha$  and  $G_\beta$ . Due to the assumed symmetry  $G_1$  and  $G_2$  are in  $R_h$ . Then eqn (3) implies that

$$\dot{V}_1 \leq -h \|x\|^2 \quad \text{for } x \in X$$

### Remarks

If  $h = 0$  in eqn (3) and  $V(x)$  is a definite function, then the stability of eqn (1) can be concluded.

The inequalities (6) which are sufficient for asymptotic stability or for instability are of course not necessary. It is easy to construct examples for which the system is stable even though the term  $g(x, t)$  is not in the region defined by these inequalities. Nevertheless, the result is general enough to provide system design criteria for a class of non-linear systems.

The region  $R_h$  is obviously dependent upon the characteristics of the term  $f(x)$ . This is consistent with the fact that the term  $f(x)$  can be designed to 'overpower' the term  $g(x, t)$ . This dependence can be used to determine the type of control  $f(x)$  which accommodates the greatest uncertainty or variation in  $g(x, t)$ .

The region  $R_h$  also depends upon the Liapunov function  $V$ , and for fixed  $f(x)$  it is desired to choose  $V$  in order to give the most suitable bounds on the function  $g(x, t)$ .

If only one component of  $g$  is non-zero, the inequalities (6) can be interpreted geometrically in the  $n + 1$  dimensional Euclidian space  $(x, z)$ . In this space there is a region,  $Z$ , made up of points which satisfy the inequalities

$$x^T g_{i\alpha} \leq z \leq x^T g_{i\beta}$$

for some  $g_{i\alpha}$  and  $g_{i\beta}$  in  $R_h$ . If the function  $z = g_i(x, t)$  is always inside the region  $Z$ , then the system is asymptotically stable (unstable). The boundary of the region  $Z$  is found by taking

$$\max_{g_{i\alpha} \in R_h} [x^T g_{i\alpha}] \quad (10)$$

and

$$\min_{g_{i\alpha} \in R_h} [x^T g_{i\alpha}]$$

for each  $x$ . From the linearity of these relations it is clear that the vectors  $g_{i\alpha}$  corresponding to the boundary of  $Z$  are on the finite or infinite boundary of  $R_h$ . The boundary of  $Z$  consists of rays from the origin and thus a map of this region may be made on the unit sphere. If the component  $g_i(x, t)$  is only a function of one of the coordinates,  $x_j$  i.e.,  $g_i(x_1, x_2, \dots, x_n, t) = g_i(0, 0, x_j, 0, t)$  then the region  $Z$  can easily be drawn as in Figure 1.

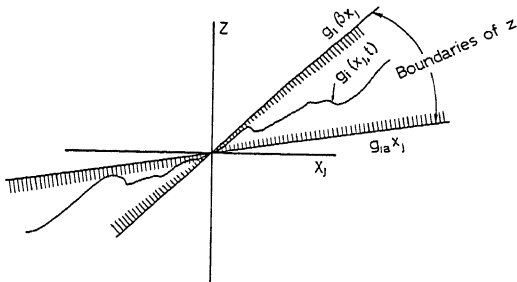


Figure 1. The region  $Z$  when the non-linearity is a function of only one variable,  $x_j$

The technique requires the use of a known Liapunov function. While there are a few methods of constructing these functions, such as the method of Zubov<sup>1</sup> and the variable gradient method due to Schultz<sup>7</sup> it may be more efficient to let  $f(x) = Ax + g_0(x)$ , where  $A$  is a constant matrix with the components

$$a_{ij} = \left. \frac{\partial f_i}{\partial x_j} \right|_{x=0, t=t_0} \quad (11)$$

The term  $g_0$  is then combined with the term  $g(x, t)$ . In this case one may choose

$$V = x^T Bx \quad \text{with} \quad B = B^T \quad (12)$$

and

$$A^T B + BA = -C \quad (13)$$

with  $C$  chosen positive definite as usual, or one may choose any positive definite  $B$ . The region  $P_h$  is determined by eqn (3) which now reads

$$x^T [(A+G)^T B + B(A+G) + hI] x \leq 0 \quad (14)$$

Thus taking

$$M = -[(A+G)^T B + B(A+G)] \quad (15)$$

The region  $P_h$  is that region in  $G$  space for which the quadratic form

$$x^T (M - hI) x > 0 \quad x \neq 0 \quad (16)$$

is positive definite. If only one row of  $G$  is non-zero,  $P_h$  is a symmetric region.

If two positive constants  $K$  and  $k$  can be chosen such that

$$K \|x\|^2 \geq V(x) \geq k \|x\|^2 \quad (17)$$

then the magnitude of the state vector is bounded by the inequality

$$\|x\| \leq \left[ \frac{K}{k} \right]^{\frac{1}{2}} e^{-\frac{h}{2K} t} \|x_0\| \quad t \geq t_0 \quad (18)$$

This inequality may be used to assure proper transient response of the controlled system. For the case  $\|x\|^2 = V$ ,  $K = k = 1$ .

When eqn (2) is asymptotically stable in  $Y$  and eqn (3) can be satisfied on the exterior of some region  $I' \subset Y$ , then eqn (6) implies that the state of the system (1) asymptotically approaches the region bounded by the minimum surface  $V = \text{constant}$  which contains the region  $I'$ .

### Application of the Technique of Linear Bounds to a Second-order System

As an example of how this theorem may be applied in the design of control systems, consider the simplified rotational dynamics of an aircraft about a single axis. The system to be considered is shown in Figure 2.

The dynamics of this system can be written in the form

$$\begin{aligned} \dot{\theta} &= \omega \\ \dot{\omega} &= f_1(\theta, \omega) + K_D f_2(\theta, u) \\ u &= -(\theta - \theta_0) - K_R \omega \end{aligned} \quad (19)$$

The control input  $u$  may be eliminated to obtain the second-order system

$$\begin{aligned} \dot{\theta} &= \omega \\ \dot{\omega} &= f_1(\theta, \omega) + K_D f_2[\theta, -(\theta - \theta_0) - K_R \omega] \end{aligned} \quad (20)$$

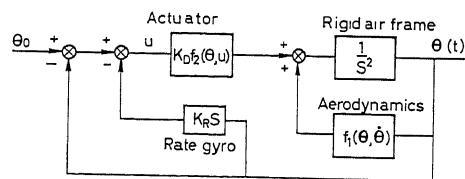


Figure 2. Simplified dynamics of aircraft control about a single axis

If an equilibrium position exists then the equation

$$f_1(\theta_e, 0) + K_D f_2(\theta_e, \theta_0 - \theta_e) = 0 \quad (21)$$

must have a solution at  $\theta = \theta_e$ . This is the requirement that a trim point exists. If there is a trim point then the variables

$$\begin{aligned} x_1 &= \theta - \theta_e \\ x_2 &= \omega \\ (\theta_e &= \text{constant}) \end{aligned} \quad (22)$$

give the deviations of the system from the trim point, and the dynamics become

$$\dot{x}_1 = x_2 \quad (23)$$

$$\dot{x}_2 = f_1(x_1 + \theta_e, x_2) + K_D f_2(x_1 + \theta_e, \theta_0 - x_1 - \theta_e - K_R x_2)$$

By choosing the numbers  $a_1$  and  $a_2$  such that  $(-a_1 x_2 - a_2 x_2)$  forms a reasonable linear approximation to the function

$$f_1 + K_D f_2$$

and taking

$$g_2(x_1, x_2, \theta_0) = f_1 + K_D f_2 + a_1 x_1 + a_2 x_2 \quad (24)$$

the dynamics now read the same as eqn (1) with

$$(x) = Ax = \begin{bmatrix} 0 & 1 \\ -a_1 & -a_2 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix}, \quad g = \begin{bmatrix} 0 \\ g_2(x_1, x_2, \theta_0) \end{bmatrix} \quad (25)$$

The Liapunov function for the equation

$$\dot{x} = Ax \quad (26)$$

is constructed as in the above remark. Taking the matrix  $C = I$  in eqn (13) the elements of  $B$  are obtained by solving the equations

$$\begin{bmatrix} 0 & 0 & -2a_1 \\ 0 & -2a_2 & 2 \\ 1 & -a_1 & -a_2 \end{bmatrix} \begin{bmatrix} b_{11} \\ b_{22} \\ b_{12} \end{bmatrix} = \begin{bmatrix} -1 \\ -1 \\ 0 \end{bmatrix} \quad (27)$$

The solution is

$$\begin{aligned} b_{11} &= \left( \frac{1}{a_2} + \frac{a_2}{a_1} \right) \frac{1}{2} + \frac{a_1}{2a_2} \\ b_{22} &= \frac{1}{2a_1 a_2} + \frac{1}{2a_2} \\ b_{12} &= \frac{1}{2a_1} \end{aligned} \quad (28)$$

the matrix  $B$  must be positive definite for stability of the linear system and this condition is satisfied for  $a_1$  and  $a_2$  positive. From the bound on the state vector in eqn (18) it is desirable to have the eigenvalues of  $B$  as small as possible for rapid transient response. The requirement that the matrix  $(\gamma I - B)$  be positive definite, assures that  $K \leq \gamma$ . This requirement places restrictions on the elements of the  $A$  matrix, and as a result the acceptable values of  $K_D$  and  $K_R$  may be obtained. The coefficients  $a_1$  and  $a_2$  are approximately given by

$$\begin{aligned} -a_1 &\approx \frac{\partial f_1}{\partial x_1} + K_D \left( \frac{\partial f_2}{\partial x_1} - \frac{\partial f_2}{\partial x_3} \right) \\ -a_2 &\approx \frac{\partial f_1}{\partial x_2} - K_D K_R \frac{\partial f_2}{\partial x_3} \end{aligned} \quad (29)$$

The coefficients  $a_1$  and  $a_2$  are chosen to be constant even though the partial derivatives may vary. The variation is part of the term  $g_2(x_1, x_2, \theta_0)$ . For this example it is assumed that  $a_1 = a_2 = 1$ . Then the matrix  $B$  is given by

$$B = \begin{bmatrix} 1.5 & 0.5 \\ 0.5 & 1.0 \end{bmatrix} \quad (30)$$

For this Liapunov function the eigenvalues are approximately  $K = 1.81$  and  $k = 0.69$ .

The vector  $g(x, t)$  has only one non-zero component and the matrix  $G$  may be taken to be

$$G = \begin{bmatrix} 0 & 0 \\ g_{21} & g_{22} \end{bmatrix} \quad (31)$$

Thus, the matrix  $M$  has the form

$$M = \begin{bmatrix} 1 - g_{21} & -g_{21} - 0.5 g_{22} \\ -g_{21} - 0.5 g_{22} & 1 - 2 g_{22} \end{bmatrix} \quad (32)$$

for  $x = x^T x$  and the matrix  $M - 1/2 I$  has the form

$$\left( M - \frac{1}{2} I \right) = \begin{bmatrix} \frac{1}{2} - g_{21} & -g_{21} - 0.5 g_{22} \\ -g_{21} - 0.5 g_{22} & \frac{1}{2} - 2 g_{22} \end{bmatrix} \quad (33)$$

The requirement that  $M - 1/2 I$  is positive definite gives the inequalities

$$\begin{aligned} \frac{1}{2} - g_{21} &> 0 \\ \frac{1}{2} - 2 g_{22} &> 0 \\ \left( g_{21} - \frac{g_{22}}{2} \right)^2 + \left( g_{22} + \frac{g_{21}}{2} \right)^2 &< \frac{1}{4} \end{aligned} \quad (34)$$

The first two inequalities are satisfied in the region

$$\begin{aligned} g_{21} &< \frac{1}{2} \\ g_{22} &< \frac{1}{4} \end{aligned} \quad (35)$$

The third inequality is a quadratic form. This equation can be simplified by a rotation of axes defined by

$$\begin{bmatrix} u \\ v \end{bmatrix} = \frac{1}{(1.25)^{1/2}} \begin{bmatrix} 1 & -0.5 \\ +0.5 & 1 \end{bmatrix} \begin{bmatrix} g_{21} \\ g_{22} \end{bmatrix} \quad (36)$$

then the quadratic form is

$$u^2 < -\frac{v}{1.118} + 0.2 \quad (37)$$

The region  $R_{1/2}$  shown in Figure 3 is bounded by this parabola. By similar computation the region  $R_0$  is obtained. Region  $Z_{1/2}$  is shown in Figure 4.

The region  $R_0$  is, as mentioned previously, only sufficient for asymptotic stability and, since the solution to the Aizerman

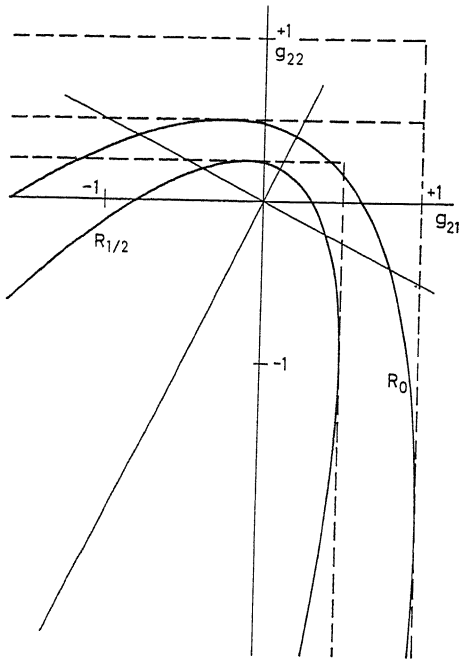


Figure 3. The  $G$ -space regions  $R_0$  and  $R_{1/2}$  for the second-order system

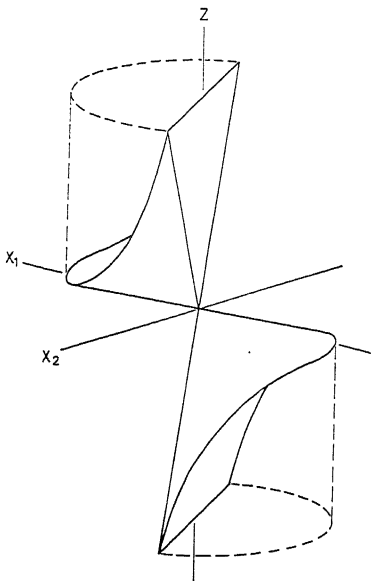


Figure 4. The region  $Z$  corresponding to  $R_{1/2}$  for the second-order system

problem for  $n = 2$  has been obtained, it is known that for  $g_{21} = 0$  ( $g_{22} = 0$ ) a sufficient condition for asymptotic stability is  $g_{22} < 1$  ( $g_{21} < 1$ ). The treatment by the method of linear bounds is more general since the non-linear terms may depend upon more than one component of the state vector. However, this generality may lead to results which are too restrictive.

It should be noted however that this set of linear bounds is much less restrictive than the bound

$$\|g\| \leq c_0 \|x\| \quad (38)$$

since  $c_0 < 1$  and this corresponds to a circle of radius 1 in the  $g_{21} g_{22}$  plane.

Using the variable gradient technique<sup>7</sup> on the equation

$$\begin{aligned} \dot{x}_1 &= x_2 \\ \dot{x}_2 &= -m_1(x_1) - m_2(x_2) \end{aligned} \quad (39)$$

one easily obtains the Liapunov function

$$V = \int_0^{x_1} m_1(\xi) d\xi + x_2^2 \quad (40)$$

with the derivative

$$\dot{V} = -2x_2 m_2(x_2) \quad (41)$$

giving the conditions

$$\begin{aligned} x_1 m_1(x_1) &> 0 \quad x_1 \neq 0 \\ x_2 m_2(x_2) &> 0 \quad x_2 \neq 0 \\ \lim_{x_1 \rightarrow \infty} \int_0^{x_1} m_1(\xi) d\xi &= \infty \end{aligned} \quad (42)$$

which are sufficient for asymptotic stability. If the function  $g_2(x_1, x_2)$  is the sum of two functions so that

$$g_2(x_1, x_2) = -n_1(x_1) - n_2(x_2) \quad (43)$$

then one obtains the conditions

$$\begin{aligned} [a_1 x_1 + n_1(x_1)] x_1 &> 0 \quad (< 0) \quad x_1 > 0 \quad (< 0) \\ [a_2 x_2 + n_2(x_2)] x_2 &> 0 \quad (< 0) \quad x_2 > 0 \quad (< 0) \end{aligned} \quad (44)$$

These relations correspond to an  $R_0$  consisting of the entire quadrant  $g_{21} < a_1$ ,  $g_{22} < a_2$  for the form of non-linearity in eqn (43).

However, because  $h = 0$  for this Liapunov function, the bound on the transient response given by eqn (18) is trivial. A bound on the response is important for control system applications.

### Application to a Third-order System

In addition to the dynamics of the airframe, the dynamics of the control actuator may be important. Consider the system in Figure 5. The dynamic equations are

$$\begin{aligned} \dot{\theta} &= \omega \\ \dot{\omega} &= a^*(\theta, u) \\ \dot{u} &= -d_1(\theta - \theta_0) - d_2\omega - d_3u \end{aligned} \quad (45)$$

The equilibrium position is given by the solution to

$$\begin{aligned} a^*(\theta_e, u_e) &= 0 \\ -d_1(\theta_e - \theta_0) - d_3u_e &= 0 \end{aligned} \quad (46)$$

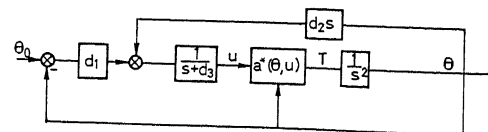


Figure 5. Third-order system in which the torque is a function  $a^*$  of the attitude and control deflection

or, combining these equations,

$$a^* \left[ \theta_e, -\frac{d_1}{d_3}(\theta_e - \theta_0) \right] = 0 \quad (47)$$

If at each instant of time this equation has the solution

$$\begin{aligned} \theta &= \theta_e \\ u &= u_e \end{aligned} \quad (48)$$

then the variables

$$\begin{aligned} x_1 &= \theta - \theta_e \\ x_2 &= \omega - \dot{\theta}_e \\ x_3 &= u - u_e \end{aligned} \quad (49)$$

define the deviations from the equilibrium position.

The deviation dynamics are given by

$$\begin{aligned} \dot{x}_1 &= x_2 \\ \dot{x}_2 &= a^*(x_1 + \theta_e, x_3 + u_e) - \dot{\theta}_e \\ \dot{x}_3 &= -d_1 x_1 - d_2 x_2 - d_3 x_3 - d_2 \dot{\theta}_e - \dot{u}_e \end{aligned} \quad (50)$$

By assuming that the equilibrium position is fixed and making a linear approximation to the function  $a^*$ , one obtains

$$\begin{bmatrix} \dot{x}_1 \\ \dot{x}_2 \\ \dot{x}_3 \end{bmatrix} = \begin{bmatrix} 0 & 1 & 0 \\ -k_1 & 0 & k_3 \\ -d_1 & -d_2 & -d_3 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} + \begin{bmatrix} 0 \\ g_2(x_1, x_3) \\ 0 \end{bmatrix} \quad (51)$$

There are six equations from eqn (13) to solve for  $B$ . Taking  $k_1 = 7.15$ ,  $k_3 = 3.58$ ,  $d_1 = 575$ ,  $d_2 = 143$ ,  $d_3 = 15$ , the matrix  $B$  (with  $C = 1$ ) is given by

$$B = \begin{bmatrix} 12.5 & 0.188 & -0.00146 \\ 0.188 & 1.4 & 0.00481 \\ -0.00146 & 0.00481 & 0.035 \end{bmatrix} \quad (52)$$

The eigenvalues of the matrix  $A$  (the 'poles' of the linear part of the system) are

$$\begin{aligned} \lambda_1 &= -4.6 \\ \lambda_{2,3} &= -5.1 \pm j21.1 \end{aligned} \quad (53)$$

and the eigenvalues of  $B$  are

$$\begin{aligned} K &= 12.55 \\ k_0 &= 1.395 \\ k &= 0.0345 \end{aligned} \quad (54)$$

The bound on the transient response in eqn (18) becomes

$$\|x\| \leq 19.1 e^{-\frac{t}{25.1}} \|x_0\| \quad (55)$$

for the Euclidian norm  $x^T x = \|x\|^2$ . This bound is quite conservative since from the solution of the linear system a more reasonable estimate is

$$\|x\| \leq 20 e^{-4t} \|x_0\| \quad (56)$$

The matrix  $G$  has only two non-zero components,  $g_{21}$  and  $g_{23}$ . The region  $R_0$  for these parameters is shown in Figure 5 and is given by the elliptic region

$$1.959 g_1^2 + 0.0018 g_1 g_2 + 1.989 g_2^2 + 0.376 g_1 + 0.0096 g_2 < 1 \quad (57)$$

The bound given by eqn (38) is quite similar to that corresponding to the region obtained above, and the bound on the transient response is too conservative to be useful. However, by using the actual minimum of  $(-\dot{V}/V)$  one may improve this estimate.

An alternative way of selecting the  $B$  matrix is as follows:

$$B = [\Phi \tilde{\Phi}^T]^{-1} \quad (58)$$

where  $\Phi$  is the matrix of eigenvectors of the matrix  $A$ . In the normal coordinate system

$$z = \Phi^{-1} x, \quad \Phi^{-1} A \Phi = \Lambda \quad (59)$$

the dynamics of the system are given by

$$\begin{aligned} \dot{z}_1 &= \lambda_1 z_1 \\ &\vdots \\ \dot{z}_n &= \lambda_n z_n \end{aligned} \quad (60)$$

A reasonable norm for the state of the system is  $\|z\| = (\bar{z}^T z)^{1/2}$ . This norm is suitable for use as a Liapunov function and has the property that

$$\frac{\dot{V}}{V} \leq 2 R_e \lambda_m < 0 \quad (\text{for stability}) \quad (61)$$

where  $\lambda_m$  is the eigenvalue with the maximum real part.

This same norm can be used in the  $x$  system of coordinates and is given by

$$\|z\|^2 = x^T B x \quad (62)$$

with  $B$  from eqn (58). For the third-order system above, this norm satisfies the following inequality

$$\|z\| \leq e^{-4t} \|z_0\| \quad (63)$$

It is suggested that this norm is more useful than the rather arbitrary norm  $x^T x$ . For  $g(x, t)$  in  $R_h$  defined using this norm, the inequality eqn (18) becomes

$$\|z\| \leq e^{-\frac{ht}{2}} \|z_0\| \quad (64)$$

The matrix  $B$  given by eqn (58) for the system of eqn (51)

$$B = \begin{bmatrix} 2.138 & 0.273 & 0.015 \\ 0.273 & 0.0617 & 0.00416 \\ 0.015 & 0.00416 & 0.00151 \end{bmatrix} \quad (65)$$

and the region  $R_{1/2}$  is also shown in Figure 6 for this case.

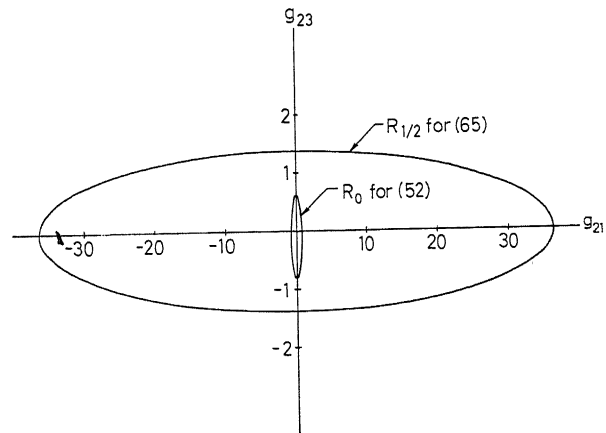


Figure 6. The regions  $R_0$  and  $R_{1/2}$  for the two different Liapunov functions for the third-order system

## Conclusion

A theorem of linear bounds is proved. This theorem may be used to obtain a class of similar systems for which satisfactory response is assured. The class found by this method is not the whole class of satisfactorily operating systems but often gives a class large enough to be of practical value. The most suitable norm for a linear system with constant coefficients is the length of the state vector in normal coordinates.

The research leading to this paper was supported in part by the United States Air Force under Contract No. AFOSR-62-68 monitored by the Air Force Office of Scientific Research of the Air Research and Development Command.

## References

- <sup>1</sup> HAHN, W. *Theorie und Anwendung der direkten Methode von Liapunov*. 1959. Berlin; Springer-Verlag

- <sup>2</sup> LIAPUNOV, M. A. Problème général de la stabilité du mouvement. *Ann. Fac. Sci. Toulouse* 9 (1907), 203
- <sup>3</sup> KALMAN, R. E. and BERTRAM, J. E. Control system analysis and design via the 'second method' of Lyapunov. Pt 1 — Continuous-time systems. *J. Basic Engng, Trans. Amer. Soc. mech. Engrs* No. 59-NAC-2
- <sup>4</sup> KALMAN, R. E. and BERTRAM, J. E. Control system analysis and design via the 'second method' of Lyapunov. Pt 2 — Discrete-time systems. *J. Basic Engng, Trans. Amer. Soc. mech. Engrs* No. 59-NAC-3
- <sup>5</sup> LETOV, A. M. *Stability in Nonlinear Control Systems*. 1961. Princeton, New Jersey; Princeton University Press
- <sup>6</sup> LA SALLE, J. and LEFSCHETZ, S. *Stability by Liapunov's Direct Method with Applications*. 1961. New York; Academic Press
- <sup>7</sup> GIBSON, J. E. and SCHULTZ, D. G. The variable gradient technique for generating Liapunov functions. *ASTIA Document No. TR-EE 62-3* (April 1962)
- <sup>8</sup> ANTOSIEWICZ, H. A. Stable systems of differential equations with integrable perturbation term. *J. Lond. Math. Soc.* 31 (1956), 208

## DISCUSSION

J. C. GILLE, *Ecole Nationale Supérieure de l'Aéronautique, 32 Boulevard Victor, Paris, France*

The second method of Liapunov is very elegant, but the application is rather difficult, since there is no general method to find the suitable Liapunov function.

Mr. Nesbit starts from the form of a standard which he introduces as a mathematical notion. It seems to me that it would be possible to introduce similar forms based on physical notions, particularly of the conservative energy in the sense of the Lagrange equations. I should like to give here the main ideas of the energy theory of stability that I suggested some time ago<sup>1</sup>. Let the non-linear equation be, for example

$$\alpha \ddot{x} + f(x, \dot{x}, t) \dot{x} + \beta x = 0$$

The conservative energy of this equation (or, more particularly, the energy of a dynamic system described by this equation) will be, after the introduction of the new variables

$$W_1 = \dot{x} \sqrt{\alpha}, \quad W_2 = x \sqrt{\beta}$$

$$E = W_1^2 + W_2^2 = \|W\|^2$$

$W_1$  and  $W_2$  may be called the plan of rectangular coordinates, in this plan the distance from the origin being a measure of energy. The system is stable if this conservative energy is diminishing in a monotonic way, i.e. if

$$\frac{dE}{dt} < 0$$

this corresponding with Liapunov's second condition.

In the case of the Hahn equation, for instance,

$$\ddot{x} + a\dot{x} + b\dot{x}^3 + x = 0$$

and

$$E = \dot{x}^2 + x^2 = \int_0^t (a + b\dot{x}^2) \dot{x}^2 dt + C$$

and thus

$$\frac{dE}{dt} = a + b\dot{x}^2 < 0$$

which means that

$$\dot{x}^2 < -\frac{a}{b}$$

(in case  $a < 0$ ,  $b > 0$ ). Based on the notion of the energy plan, we conclude that there is stability inside of the circle centred at the origin, the square of the radius being equal to  $-a/b$  (Figure A). The notion of the energy plan, the conservative energy plan and

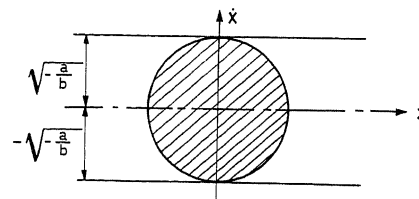


Figure A

of the conservative energy permits, as a consequence, to find not only the condition of unlimited stability, but also of the total stability within a certain domain as compared with the initial conditions  $x_0, \dot{x}_0$

$$x_0^2 + \dot{x}_0^2 < -\frac{a}{b}$$

We hope it will be possible to extend the energy theory of stability to systems of a higher order, but for this some new notions should be introduced, such as conservative and dissipative coefficients, conservative energy and the associated conservative system.

We also hope to present a more complete theory for the next I.F.A.C. Congress

## Reference

- <sup>1</sup> GILLE, J. C. and WEGRZYN, S. Associated conservative systems and non-linear stability. *Automatisme VII(S)* (1962)

R. A. NESBIT, in reply

The use of energy and momentum integrals for Liapunov functions is consistent with the historical motivations of the theory. In active systems the possibility of making the system unstable (making the 'energy' increase) gives stability its practical significance. I am interested in the general problem of determining Liapunov functions and look forward to the extension of the energy method.

M. MANSOUR, *Institut für Automatik und Industrielle Elektronik ETH, Zurich, Switzerland*

In order to prove the theorem of linear bounds the author selected a subset  $R_h$  from the set  $P_h$  in  $G$  space for which the Sylvester inequalities are valid or, more generally, for which the derivative of Liapunov function are sign definite. It is not easy to select this subset, especially when the system is of high dimension or if the original system is already non-linear. Sometimes it is not possible to find a symmetric subset in the  $G$  space because it does not exist, e.g. a system where the equations are

$$\dot{z}_i = -\lambda_i z_i + f(\sigma) \quad \sigma = \sum \alpha_i z_i \quad i = 1, \dots, n$$

This can be put in the vector form  $\dot{Z} = AZ + g(z)$  and the non-linear  $f(0)$  represents a non-linear element in series with a linear element in a closed loop. If we succeed in selecting a symmetric subset, e.g. by taking all the rows of  $G$  equal to zero except one for the sake of simplifying the calculations, this will impose severe restrictions on the non-linearity, especially if the system is of high dimension so that the method may not yield practical results.

R. A. NESBIT, *in reply*

A symmetric set exists since  $G = 0$  is one element of the set. The following procedure can be used computationally to determine a symmetric set of matrices: Choose 'directions' in  $G$  space by specifying possible rows for the matrices  $G_i$ . Form a set  $L$  of matrices by choosing each combination of the specified rows. Multiply each member of the set  $L$  by a scalar,  $k_i$ . The derivative  $\dot{V}_1$  is now a function of the scalar  $k_i$ , and  $\dot{V}_1$  is negative definite for  $k_i = 0$ . The interval  $-a_i \leq k_i \leq b_i$  for which  $\dot{V}_1$  remains negative definite must be determined from eqn (3). From the intersection of all the intervals one can obtain an interval  $-a \leq k \leq b$ .

The set obtained by multiplying each member of  $L$  by any  $k$ :  $-a \leq k \leq b$  is a symmetric subset  $R_n$ . This method is clearly not practical when the number of possible rows is so great that the combinatorial number of elements in  $L$  makes computational time prohibitive. Furthermore the results will be disappointing if the interval  $[-a, b]$  is very short. In the example

$$\dot{z}_i = -\lambda_i z_i + f(\sigma) \quad (i = 1, 2 \dots n)$$

$$\sigma = \sum_{i=1}^n \alpha_i z_i$$

The choice of a symmetric subset is especially easy since all the rows of  $G$  can be taken the same. Further simplification results in this case because the argument of the non-linearity is a linear combination of state variables and thus only one direction in  $G$  space is of interest. Instead of a combinational analysis, only one matrix  $G$  need be considered:

$$G = \begin{bmatrix} \alpha_1 & \alpha_2 & \dots & \alpha_n \\ \alpha_1 & \alpha_2 & \dots & \alpha_n \\ \alpha_1 & \alpha_2 & \dots & \alpha_n \end{bmatrix}$$

From this one matrix, one obtains bounds of the form  $-\alpha\sigma \leq f(\sigma) \leq b\sigma$ . It is important to recognize that it is not necessary to take all rows equal to zero, except one, to form a symmetric set of matrices.

M. R. PATEL, *Cambridge University, Cambridge, England*

I would like to present a small extension to Schultz's method of constructing Liapunov functions via  $\nabla V$ . The requirement for  $V$  to be single valued is that  $\nabla \times \nabla V = 0$ . Another way of putting the same thing is that the second-order cross derivatives of  $V$  be equal, i.e.

$$\frac{\partial^2 V}{\partial x_i \partial x_j} = \frac{\partial^2 V}{\partial x_j \partial x_i}, \quad i \neq j$$

That gives us

$$\frac{n(n-1)}{2}$$

number of equations to aid the choosing of Liapunov function  $V$ . This can be carried further to higher order cross-derivatives, and say that

$$\frac{\partial^3 V}{\partial x_i \partial x_j \partial x_k} = \frac{\partial^3 V}{\partial x_j \partial x_k \partial x_i} = \frac{\partial^3 V}{\partial x_k \partial x_i \partial x_j}$$

This gives

$$\frac{n(n-1)-2}{2}$$

more equations to aid the choosing of the Liapunov functions. However neat the method may be, solving several partial differential equations simultaneously will be no easy task. It is interesting to note, that nearly all systematized methods of constructing Liapunov functions come back to solving partial differential equations such as Zubov's method, Schultz's method and Mason's extension of Zubov's method.

P. C. PARKS, *Department of Aeronautics, The University, Southampton, England*

The author is concerned with aeronautical stability problems. I was concerned with the solution, by the second method of Liapunov, of an interesting missile stability problem<sup>1</sup> involving a system of linear equations with periodic coefficients. With a quadratic form as Liapunov function, it was possible to obtain a useful sufficient stability criterion and, within the stability region so found, estimates of the decay of transient motion along the lines of eqn (18), due originally to N. G. Chetaev.

#### Reference

- 1 PARKS, P. C. Pitch-yaw stability of a missile oscillating in roll via the second method of Liapunov. *J. Aerospace Sci.* 29, No. 7 (1962)

R. A. NESBIT, *in reply*

I thank Mr. Parks for giving an interesting application of the method to an equation with time-varying coefficients.

G. P. SZEGÖ, *Research Institute for Advanced Studies (R.I.A.S.), Baltimore 12, Maryland, U.S.A.*

I congratulate Dr. Nesbit on his very fine paper, and make the following comment. The ultimate aim of the procedure presented is of course to find the best quadratic form which will yield the maximum allowable range of variation of the non-linear perturbation  $g(x, t)$ . In the case in which the non-linear perturbation depends only upon one of the components of the vector  $x$ , say  $x_i$ , the characterization of the 'best quadratic form' is quite simple: it is the quadratic form which allows the widest possible wedge in which  $g(x_i)$  has to lie. However, when  $g(x, t)$  depends upon more than one component of the state vector  $x$ , the definition of 'best quadratic form' does not seem to be a trivial problem and depends upon the particular system under investigation.

What particular figure of merit for the definition of 'best quadratic form' did the author use in this work? Would he care to comment on this problem in general?

R. A. NESBIT, *in reply*

This important problem of optimizing the choice of Liapunov function is not directly considered in the paper. As with all problems of optimization, choice of criterion is antecedent to computational solution. The comparative definition of value used by the author was that the union of two 'wedges' is better than one 'wedge'.

Value is subjective, and the definition of 'best' Liapunov function must depend upon the individual situation. However, in the case that the relative magnitudes of the bounds on various components of  $g$  can be specified, there is still no convenient method for determining the 'optimum' Liapunov function.

# On the Estimation of the Decaying Time

L. HWANG

## Summary

In this paper Liapunov's second method is applied to the stability for certain functions taking on given values, and some formulae, to estimate the decaying time in the general case, are obtained. The linear system with quasi-constant coefficients, and the quasi-reducible system, are studied in detail.

## Sommaire

La seconde méthode de Liapunov est appliquée à l'étude de la stabilité de certaines fonctions avec valeurs données. On a obtenu des formules pour l'estimation du temps d'amortissement dans le cas général. Le système linéaire à coefficients quasi-constants et le système quasi-réductible sont étudiés en détails.

## Zusammenfassung

Dieser Aufsatz behandelt mittels der 2. Methode von Liapunov das Stabilitätsproblem für bestimmte Funktionen, die gegebene Werte annehmen. Daraus ergeben sich allgemeine Beziehungen zur Abschätzung der Abklingzeit. Außerdem werden lineare Systeme mit quasi-konstanten Koeffizienten und quasi-reduzierbare Systeme im einzelnen untersucht.

The problem of stability is one of the basic problems in the proper operation of any dynamic system. After Liapunov's brilliant work<sup>1</sup>, very great effort has been expended on its theory in the last twenty years<sup>2-8</sup>. In the monograph by Zubov<sup>4</sup>, the stability of the invariant sets of the abstract dynamic systems in the metric space are treated in a very general sense.

Consider the differential system

$$\dot{x}_s = X_s(x_1, \dots, x_n, t) \quad (s=1, 2, \dots, n) \quad (1)$$

It is often necessary to study the stability problem for the given values  $\phi_1^0, \dots, \phi_K^0$  of the coordinate functions

$$\phi_i = \phi_i(x_1, \dots, x_n) \quad (i=1, 2, \dots, K) \quad (2)$$

Without loss of generality,  $\phi_1^0 = 0, \dots, \phi_K^0 = 0$  may be taken. Thus, the standard working state of the system is given by the equation

$$\phi_i(x_1, \dots, x_n) = 0 \quad (i=1, 2, \dots, K) \quad (3)$$

Let the set of points defined by (3) constitute an  $(n-k)$  manifold  $\mathcal{F}(n-k)$ . By means of (3), some particular and interesting motions of system (1) can generally be described (for example, the self-excited oscillations or the motions which demand their characteristic functions to take on given values). Here the generalized stability differs from the Liapunov stability in that the unperturbed motion is no longer a particular motion (e.g. trivial solution) but its  $K$  coordinate functions take on given values. In general, (3) represents a class of motions and constitutes a manifold in the phase-space. From the research point of view, the coordinate functions (2) are of more interest than the coordinates  $x_1, \dots, x_n$  themselves.

Obviously, when  $\phi_i = x_i, i=1, \dots, k, k < n$ , the stability of the partial coordinates is obtained, and when  $\phi_i = x_i, i=1, \dots, k, k=n$ , it agrees with the stability in Liapunov's sense.

In papers by Liapunov<sup>1</sup> and Rumyantsev<sup>8</sup> the stability of the partial coordinates and the stability for the given values of the functions are discussed. They require the absolute values of the initial perturbations  $x_1^0, \dots, x_n^0$  to be sufficiently small and thus essentially the unperturbed motion was supposed to be a point in the phase space. The approach in this paper differs from theirs, and it will be explained clearly below.

The set of points which satisfy the inequality

$$\sum_{i=1}^K \phi_i^2(x_1, \dots, x_n) \leq H^2 \quad (4)$$

is called the  $H$ -neighbourhood of  $\mathcal{F}(n-k)$  and is written as  $\mathcal{F}(n-k)(H)$ . It is assumed tacitly that through each point of  $\mathcal{F}(n-k)(H)$  there exists a unique solution of the system (1).

Of course, it is necessary that the standard working state (3) of the system has a certain upholding ability, which means that the functions

$$\Phi_i(x_1, \dots, x_n, t) = \text{grad } \phi_i X \quad (i=1, \dots, K) \quad (5)$$

should satisfy the conditions

$$\Phi_i(x_1, \dots, x_n, t) \equiv 0 \text{ as } \{x\}_n \in \mathcal{F}(n-k) \quad (6)$$

or equivalently,  $\mathcal{F}(n-k)$  is an invariant set of the system (1), where  $\{x\}_n$  represents the  $n$ -dimensional vector.

**Definition 1.** The system (1) is said to be stable with respect to the functions (2), taking on zeros (3) whenever, given any  $\varepsilon > 0$ , there is a  $\delta(t_0, \varepsilon) > 0$ , such that, for all trajectories  $x(t)$  with initial values satisfying

$$\{x(t_0)\}_n = \{x^0\}_n \in \mathcal{F}(n-k)(\delta), \quad t_0 > 0 \quad (7)$$

one has

$$\{x(t)\}_n \in \mathcal{F}(n-k)(\varepsilon) \quad (8)$$

for all  $t \geq t_0$ .

**Definition 2.** The system (1) is said to be asymptotically stable with respect to the functions (2) taking on zeros (3) if

(a) Definition (1) holds,

$$(b) \quad \lim_{t \rightarrow \infty} \Phi^2 = \lim_{t \rightarrow \infty} \sum_{i=1}^K \phi_i^2[x(t), \dots, x_n(t)] = 0 \quad (9)$$

i.e. for any given  $\eta > 0$  there is a positive number  $T = T(\eta, t_0, x^0)$  such that

$$\{x(t)\}_n \in \mathcal{F}(n-k)(\eta) \text{ as } t \geq t_0 + T \quad (10)$$

**Definition 3.** The system (1) is said to be equi-asymptotically stable with respect to the functions (2), taking on zeros (3) if

(a) Definition 2 holds,



(b) there exists  $H > 0$ , and  $T = T(\eta)$ , such that for all trajectories  $\{x(t)\}_n$  with initial values satisfying

$$\{x(t_0)\}_n \in \mathcal{F}(n-k)(H) \quad t_0 \geq 0 \quad (11)$$

then (10) holds.

If the initial conditions were subjected to the following restrictions

$$\{x(t_0)\}_n \in G^0 \quad t_0 > 0 \quad (12)$$

then the above-mentioned stability, asymptotic stability and equi-asymptotic stability are said to be stable, asymptotically stable and equi-asymptotically stable under condition (12) respectively.

In the sequel, the function  $V(x_1, \dots, x_n, t)$  is called the Liapunov function with respect to functions  $\phi_1, \dots, \phi_k$  if

$$V(x_1, \dots, x_n, t) \equiv 0 \text{ as } \{x\}_n \in \mathcal{F}(n-k) \quad (13)$$

and  $V$  is assumed to have continuous partial derivations.

**Definition 4.** The function  $V(x_1, \dots, x_n, t)$  is said to be positive (negative) semi-definite with respect to (2) if

(a) (13) holds,

(b)  $V \geq 0$  [ $V \leq 0$ ] in  $\mathcal{F}(n-k)(H)$ .

**Definition 5.** The function  $V(x_1, \dots, x_n, t)$  is said to be positive (negative) definite with respect to (2) if

(a) Definition 4 holds,

(b) there is a positive function  $W_1(y_1, \dots, y_k)$  such that, in  $\mathcal{F}(n-k)(H)$ ,

$$V(x_1, \dots, x_n, t) \geq W_1[\phi_1(x_1, \dots, x_n), \dots, \phi_k(x_1, \dots, x_n)] \quad (14)$$

$$\{V(x_1, \dots, x_n, t) \leq -W_1[\phi_1(x_1, \dots, x_n), \dots, \phi_k(x_1, \dots, x_n)]\}$$

**Definition 6.** The function  $V$  is said to be uniformly small, if for any given  $\varepsilon > 0$ , there is  $\delta(t) > 0$  such that the conditions  $t > 0$  and  $\{x\}_n \in \mathcal{F}(n-k)(\delta)$  imply  $V \leq \varepsilon$ .

**Definition 7.** The function  $V(x_1, \dots, x_n, t)$  is said to have infinitely small upper bound with respect to (2) if there is a continuous function  $W_2(y_1, \dots, y_k)$  such that

(a)  $W_2(0, \dots, 0) = 0$ ,

(b) in  $\mathcal{F}(n-k)(H)$ ,

$$W_2[\phi_1(x_1, \dots, x_n), \dots, \phi_k(x_1, \dots, x_n)] \geq V(x_1, \dots, x_n, t) \quad (15)$$

**Definition 8.** The function  $V(x_1, \dots, x_n, t)$  is said to have the property  $A$ , if there are two positive continuous functions  $W_1(S)$  and  $W_2(S)$  such that

$$(a) \quad W_1(0) = W_a(0) = 0, \quad W_1(\infty) = W_a(\infty) = +\infty \quad (16)$$

$$(b) \quad W_2(\|\phi\|_K) \geq V(x_1, \dots, x_n, t) \geq W_1(\|\phi\|_K) \quad (17)$$

Parallel to Definitions 1 and 3, one has the fundamental theorems shown in the following section.

### The Fundamental Theorems

(A) For the system (1), if there is a Liapunov function  $V(x_1, \dots, x_n, t)$  such that

(a)  $V$  satisfies Definition 5 and it is positive definite with respect to (2),

(b)  $V$  satisfies Definition 7,

(c) the total derivative

$$\left. \frac{dV}{dt} \right|_{1.1} = \frac{\partial V}{\partial t} + \text{grad } V \cdot X \quad (18)$$

is negative definite with respect to (2), then the system (1) satisfies Definition 3.

(B) If the system (1) satisfies Definition 3, and the rank of matrix

$$\frac{D(\phi_1, \dots, \phi_K)}{D(x_1, \dots, x_n)} = \begin{pmatrix} \frac{\partial \phi_1}{\partial x_1}, \dots, \frac{\partial \phi_1}{\partial x_n} \\ \dots \\ \frac{\partial \phi_K}{\partial x_1}, \dots, \frac{\partial \phi_K}{\partial x_n} \end{pmatrix} \quad (19)$$

is  $K$ , and the functions  $\Phi_i$  ( $i = 1, \dots, K$ ) defined by (5) are uniformly bounded in  $\mathcal{F}(n-k)(H)$ , then there is a function  $V$  which satisfies all conditions in (A). The proof of this theorem is given in the Appendix. It is not difficult to prove the following corollaries.

**Corollary 1.** If  $V$  satisfies Definitions 5 and 6, and  $dV/dt|_{(1)}$  is negative semi-definite with respect to (2), then the system (1) satisfies Definition 1.

**Corollary 2.** If  $V$  satisfies Definition 8, and  $dV/dt|_{(1)}$  is negative definite with respect to (2) then (9) holds for any  $t_0 > 0$  and any  $x^0$  in the space.

Let the set of position points at time of the motions which take on the initial positions in  $G^0$  be written as  $G^{(t)}$ .

**Corollary 3.** If  $V$  satisfies (A), Corollary 1 or Corollary 2 in  $G^{(t)} \cap \mathcal{F}(n-k)(H)$ , then the system (1) is stable, equi-asymptotically stable, or asymptotically stable in the whole under condition (12) respectively.

**Corollary 4.** If the system (1) satisfies Definition 3 under condition (12) and the rank of matrix (19) is  $K$  in the neighbourhood  $\mathcal{F}(n-k) \cap G^{(t)}$  and  $\Phi_i$  ( $i = 1, \dots, K$ ) are uniformly bounded in  $G^{(t)} \cap \mathcal{F}(n-k)(H)$ , then there is a function  $V$  which satisfies the conditions in Corollary 3.

**Example—**Consider the system

$$\begin{cases} \dot{x} = ay - cx(bx^2 + ay^2) \sin \frac{1}{bx^2 + ay^2} \\ \dot{y} = -bx - cy(bx^2 + ay^2) \sin \frac{1}{bx^2 + ay^2} \end{cases} \quad (20)$$

$$(a \cdot b \cdot c > 0)$$

Obviously, if one takes  $\phi = bx^2 + ay^2$  then  $\phi = 1/k\pi$  ( $k = 1, 2, \dots$ ) are the invariant sets of (20), they are closed orbits. By means of the Liapunov functions  $V = \frac{1}{2}(\phi - 1/k\pi)^2$  with respect to  $\phi - 1/k\pi$ , the following statements can be proved:

(a) In the exterior of the ellipse  $\phi = 1/\pi$ , there is no closed orbit;

(b) in the interior of the ellipse  $\phi = 1/\pi$ , there are infinitely many closed orbits;

(c) the closed orbit is asymptotically stable when  $K$  is even and it is unstable when  $K$  is odd;

(d) the origin  $x = y = 0$  is a singular point of (20) and it is stable. In any of its neighbourhood, there are infinitely many closed orbits, and hence the origin is not asymptotically stable

In the control or the dynamic systems, it is often necessary to estimate the decaying time of perturbations for the standard working state. In this paper the problem of estimating the decaying time is considered. In the sequel it is assumed that system (1) is equi-asymptotically stable with respect to (2), and the following discussions are valid in certain attractive regions of  $\mathcal{F}(n-k)$ .

Let  $V$  be a Liapunov function of (1) which satisfies the conditions of the fundamental theorem (A). In the general case, there are two positive definite functions  $W_1(y_1, \dots, y_k)$  and  $W_2(y_1, \dots, y_k)$  such that

$$\begin{aligned} W_2[\phi_1(x), \dots, \phi_k(x)] &\geq V(x_1, \dots, x_n, t) \\ &\geq W_1[\phi_1(x), \dots, \phi_k(x)] \end{aligned} \quad (21)$$

Besides, it is assumed that there are two functions  $f_1(s)$  and  $f_2(s)$ , such that in  $\mathcal{F}(n-k)(H)$  the inequalities

$$f_1(v) \leq \frac{dv}{dt} \leq f_2(v) \quad (22)$$

hold. Furthermore, from (21) one has, in general,

$$\begin{aligned} \{x_n\} \in \{W_2 \leq V_0\} &\text{ implies } \{x\}_n \in \{V \leq V_0\} \\ \{x_n\} \in \{V \leq \varepsilon\} &\text{ implies } \{x\}_n \in \{W_1 \leq \varepsilon\} \end{aligned} \quad (23)$$

where  $V_0$  and  $\varepsilon$  are given position numbers. Denote

$$T_1 = - \int_{\varepsilon}^{V_0} \frac{d\lambda}{f_1(\lambda)} \quad T_2 = - \int_{\varepsilon}^{V_0} \frac{d\lambda}{f_2(\lambda)} \quad (24)$$

Then, the following theorem estimates the decaying time.

**Theorem 1.** The decaying time  $T$  of the motion of the system (1) from an initial point in the region

$$W_2[\phi_1(x), \dots, \phi_k(x)] \leq V_0 \quad (25)$$

to a point in the region

$$W_1[\phi_1(x), \dots, \phi_k(x)] \leq \varepsilon \quad (26)$$

satisfies the inequality

$$T \leq T_2 \quad (27)$$

The decaying time  $T$  from an initial point in the region

$$W_2[\phi_1(x), \dots, \phi_k(x)] \geq V_0 \quad (28)$$

to a point in the region (26) satisfies

$$T \geq T_1 \quad (29)$$

Let  $M_2(R)$  be the maximum value of  $W_2$  on the boundary

$$\|\phi\|_K = R \text{ of } \mathcal{F}(n-k)(R)$$

and let  $m_1(\gamma)$  be the minimum value on the boundary  $\|\phi\|_K = \gamma$  of  $\mathcal{F}(n-k)(\gamma)$ . Again denoting

$$T_1 = \int_{m_1(\gamma)}^{M_2(R)} \frac{d\lambda}{f_1(\lambda)} \quad T_2 = \int_{m_1(\gamma)}^{M_2(R)} \frac{d\lambda}{f_2(\lambda)} \quad (30)$$

the following theorem is obtained.

**Theorem 2.** The decaying time  $T$  of the motion of the system (1) from an initial point in the region  $\mathcal{F}(n-k)(R)$  to a point in the region  $\mathcal{F}(n-k)(\gamma)$  satisfies (27), and the decaying time  $T$  of the motion of the system (1) from an initial point in the region  $\|\phi\| \geq R$  to a point in the region  $\mathcal{F}(n-k)(\gamma)$  satisfies (29), where  $T_1, T_2$  are defined by (30).

By taking

$$f_1(v) = -\alpha v \quad f_2(v) = -\beta v \quad (\alpha > \beta) \quad (31)$$

one has

$$T_1 = \frac{1}{\alpha} \log \frac{M_2(R)}{m_1(\gamma)} \quad T_2 = \frac{1}{\beta} \log \frac{M_2(R)}{m_1(\gamma)} \quad (32)$$

Particularly, when  $\phi_i = x_i, i = 1, \dots, k < n$  one obtains the formulae to estimate the decaying time for partial coordinates, and when  $\phi_i = x_i, i = 1, \dots, n$ , then one obtains the formulae to estimate the decaying time for total coordinates (9-11).

The above method is used to solve the following example.

**Example—**Consider an autonomous system

$$\begin{aligned} \dot{x} &= x - a^2 x^3 - b^2 x y^2 \\ \dot{y} &= y - a^2 y x^2 - b^2 y^3 \end{aligned} \quad (a > b) \quad (33)$$

and its unique closed orbit

$$\phi = a^2 x^2 + b^2 y^2 - 1 = 0 \quad (34)$$

If one selects the Liapunov function with respect to  $\phi$  to be

$$V = (a^2 x^2 + b^2 y^2 - 1)^2 \quad (35)$$

then it may be asserted that:

(a) the system is asymptotically stable with respect to  $\phi = 0$ ;

(b) the decaying time  $T$  of the motion of the system from an initial point in the region  $|\phi| \leq \phi_0$  to a point in the region  $|\phi| \leq \varepsilon$  satisfies  $T \leq a^2 \log(1 + \varepsilon) \phi_0 / (1 + \phi_0) \varepsilon$ ;

(c) the decaying time  $T$  of the motion from an initial point in the region  $|\phi| \leq \phi_0$  to a point in the region  $|\phi| \leq \varepsilon$  satisfies  $T \geq b^2 \log(1 + \varepsilon) \phi_0 / (1 + \phi_0) \varepsilon$ .

#### On the Estimation of Decaying Time for a Linear System with Quasi-constant Coefficients

In the study of a practical dynamic system, one usually takes the linear system with constant coefficients as its first approximation. In general, the frequency method may be applied to estimate the time of transient process for the control system with constant coefficients. However, this method is only applicable to the case of single output under specific initial conditions. In addition, the method is not rigorous. This paper gives the formulae to estimate the decaying time in the general case, and the method is rigorous.

A large amount of work<sup>9-12</sup> is devoted to the estimation of decaying time for the asymptotically stable system

$$\dot{x}_S = p_{S1}x_1 + \dots + p_{SN}x_N, \quad S = 1, \dots, N \quad (36)$$

where the coefficients  $p_{ij}$  are constants. These results may be summarized as the following. For any given positive definite quadratic form

$$U = x' U x \quad (37)$$

there is a positive definite quadratic form

$$V = x' V x \quad (38)$$

such that

$$\left. \frac{dV}{dt} \right|_{(36)} = -U \quad (39)$$

If  $M_1$  and  $m_1$  are, respectively, the maximum and the minimum eigenvalues of the matrix  $V$ , and  $M$  and  $m$  are the maximum and the minimum eigenvalues of the matrix  $U$ , then the following results are obtained.

**Theorem 3.** The decaying time  $T$  of the motion of the system (36) from an initial point in

$$\sum_{s=1}^N x_s^2 = R^2$$

to a point in the region

$$\sum_{s=1}^N x_s^2 \leq r^2$$

satisfies the inequalities

$$\frac{m_1}{M_2} \log \frac{M_1 R^2}{m_1 r^2} \leq T \leq \frac{M_1}{m_2} \log \frac{M_1 R^2}{m_1 r^2} \quad (40)$$

In practice, it is of interest to select a suitable Liapunov function  $V$ , such that for the given system (36) the range defined by (40) is as accurate as it can be. It is very difficult to answer the above question in the general case. But if the system (36) is normal and the elementary divisors of the coefficient matrix  $P$  are all simple, it may be proved that when

$$V = \sum_{s=1}^N x_s^2$$

the equalities in (40) may be realized (i.e. the estimation is accurate).

Let the normal transformation be

$$y = Cx \quad (41)$$

where  $C$  is a matrix with real coefficients, and the system (36) is reduced to the normal system

$$\dot{y} = Jy \quad (42)$$

where

$$J = \begin{pmatrix} -\alpha_1 & & & 0 \\ & -\alpha_K & & \\ & & -\beta_1 - \omega_1 & \\ & & \omega_1 - \beta_1 & \\ 0 & & & -\omega_l \\ & & & \omega_l - \beta_l \end{pmatrix} \quad (43)$$

$$V = x' C' C x \quad (44)$$

may be taken as a Liapunov function of the system (36). By means of (42) the following results may be proved.

**Theorem 4.** The decaying time  $T$  of the motion of the system (36), from an initial point in the  $(n-1)$ -dimensional ellipsoid  $V = V_0$  to a point in the ellipsoid  $V = \varepsilon$ , satisfies

$$\frac{1}{2u} \log \frac{V_0}{\varepsilon} \leq T \leq \frac{1}{2v} \log \frac{V_0}{\varepsilon} \quad (45)$$

where  $u = \max(\alpha_i, \beta_j)$ ,  $v = \min(\alpha_i, \beta_j)$ . It is easy to select the initial points such that the equalities in (45) hold (i.e. this estimation is accurate).

In the following, the general formulae to estimate the decaying time is given.

All roots of the characteristic equation  $D(\lambda) = \det(R - \lambda I) = 0$  are assumed to have negative real parts. Let  $\lambda_i, \dots, \lambda_l$  be negative real roots, written as  $\lambda_i = -\alpha_i$  ( $i = 1, \dots, l$ ), and let  $\lambda_{l+1}, \dots, \lambda_n$  be the remaining roots, written as  $\beta_s \pm \omega_s i$  ( $s = 1, \dots, n-l/2 = k$ ), and the order of the corresponding elementary divisors be  $n_1, \dots, n_k$ .

It is known that there is a non-singular linear transformation  $y = Cx$  to reduce the system (36) to the normal form (42), in which

$$J = \begin{pmatrix} M_1 & & 0 \\ & \ddots & \\ & & M_l \\ & & & N_1 \\ & & & & \ddots \\ & & & & & N_k \end{pmatrix} \quad (46)$$

$$M_i = \begin{pmatrix} -\alpha_i & 1 & 0 & \dots & 0 \\ 0 & -\alpha_i & 1 & \dots & 0 \\ & & \ddots & \ddots & \\ 0 & & & 1 & \\ & & & & -\alpha_i \end{pmatrix} N_j = \begin{pmatrix} -\beta_j - \omega_j & 1 & 0 & \dots & 0 \\ \omega_j - \beta_j & 0 & 1 & \dots & 0 \\ & & \dots & & \\ 0 & 0 & \dots & -\beta_j - \omega_j & \\ 0 & 0 & \dots & \omega_j - \beta_j & \end{pmatrix} \quad (47)$$

If one writes the  $m_i \times m_i$  matrix

$$a^{(m_i)} = \begin{pmatrix} \frac{1}{\alpha_i} & \frac{1}{2\alpha_i^2} & \frac{1}{4\alpha_i^3} \dots \\ \frac{1}{2\alpha_i^2} & \frac{1}{\alpha_i} \left[ 1 + \frac{1}{2\alpha_i^2} \right] \dots \\ \frac{1}{4\alpha_i^3} \dots \end{pmatrix} \quad (48)$$

as  $a_{s\sigma}^{(m_i)}$ , this is constructed according to the following rule:

(a) when  $s = \sigma$ ,  $a_{s\sigma}^{(m_i)}$  is equal to  $(1/\alpha_i) (1 + a_{s-1, \sigma}^{(m_i)})$ , and let

$$a_{11}^{(m_i)} = \frac{1}{\alpha_i};$$

(b) when

$$s > \sigma, a_{s\sigma}^{(m_i)} = \frac{1}{2\alpha_i} [a_{s, \sigma-1}^{(m_i)} + a_{s-1, \sigma}^{(m_i)}];$$

(c)

$$a_{s\sigma}^{(m_i)} = a_{\sigma s}^{(m_i)}$$

Thus the matrix is completely defined through the eigenvalues  $-\alpha_i$  and the order of its elementary divisor. The maximum eigenvalue of the matrix  $a^{(m_i)}$  is assumed to be  $v_i$

$$\begin{cases} \text{when } m_i = 1, v_i = \frac{1}{\alpha_i} \\ \text{when } m_i = 2, v_i = \frac{1}{\alpha_i} + \frac{1}{4\alpha_i^2} \left[ \frac{1}{\alpha_i} + \left( 4 + \frac{1}{\alpha_i^2} \right)^{\frac{1}{2}} \right] \end{cases} \quad (49)$$

Following the method of construction of the matrix  $a^{(m)}$ , the  $2n_i \times 2n_i$  matrix  $d^{(2n_i)}$  may be constructed in the following manner

$$(a) \quad d_{2i-1, 2j}^{(2n_i)} = d_{2i, 2j-1}^{(2n_i)} = 0, \quad i, j = 1, \dots, n_i$$

$$(b) \quad d_{2i-1, 2j-1}^{(2n_i)} = d_{2j-1, 2i-1}^{(2n_i)} = a_{ij}^{(n_i)} \quad i, j = 1, \dots, n_i$$

(c) to replace  $\alpha_i$  by  $\beta_i$  in the matrix  $a^{(n_i)}$ .

For example  $n_i = 2$ , one has

$$a^{(4)} = \begin{pmatrix} a_{11}^{(2)} & 0 & a_{12}^{(2)} & 0 \\ 0 & a_{11}^{(2)} & 0 & a_{12}^{(2)} \\ a_{21}^{(2)} & 0 & a_{22}^{(2)} & 0 \\ 0 & a_{21}^{(2)} & 0 & a_{22}^{(2)} \end{pmatrix}$$

Obviously, the formula for the maximum eigenvalue of  $d^{(2n_i)}$  is the same as that of  $a^{(m)}$  in which  $\alpha_i$  are replaced by  $\beta_i$ . Consider the Liapunov function for system (36) to be

$$V = x' C' A C x \quad (50)$$

where  $C$  is the normal transformation matrix, and

$$A = \begin{pmatrix} a^{(m_1)} & & & 0 \\ & a^{(m_1)} & & \\ & & d^{(2n_1)} & \\ 0 & & & d^{(2n_k)} \end{pmatrix} \quad (51)$$

It is not difficult to prove that  $V$  satisfies

$$\left. \frac{dV}{dt} \right|_{2,1} \leq -\frac{2}{v} V \quad (52)$$

where  $v$  is the maximum eigenvalue of the matrix  $A$ , and it can be calculated by the aforesaid method. When  $m_i = 1$  or  $m_i = 2$  it can be calculated through (49). If the maximum and the minimum eigenvalues of the symmetric matrix  $C'AC$  are assumed to be  $M$  and  $m$  respectively, then the following theorem is obtained.

**Theorem 5.** The decaying time  $T$  of the motion of the system (36) from an initial point in the  $(N-1)$ -dimensional ellipsoid  $V = V_0$  to a point in the  $(N-1)$ -dimensional ellipsoid  $V = \varepsilon$  satisfies

$$T \leq \frac{v}{2} \log \frac{V_0}{\varepsilon}$$

The decaying time  $T$  of the motion of the system (36) from an initial point in the sphere

$$\sum_{s=1}^N x_s^2 = R^2$$

to a point in the sphere

$$\sum_{s=1}^N x_s^2 = r^2$$

satisfies

$$T \leq v \log \frac{MR^2}{mr^2}$$

Moreover, the system

$$\dot{x} = px + X(x, t) \quad (53)$$

is considered, where  $X$  is a vector function which contains the non-linear terms and the unknown components. If one constructs a Liapunov's function (50) of its principal linear system

$$\dot{x} = px \quad (54)$$

and if one assumes  $X$  to satisfy the inequality

$$|\text{grad } V \cdot X| < bx' C' C x, \quad (b < 2) \quad (55)$$

then the following results are obtained.

**Theorem 6.** The decaying time  $T$  of the motion of the system (53) from an initial point in the sphere

$$\sum_{s=1}^N x_s^2 = R^2$$

to a point in the sphere

$$\sum_{s=1}^N x_s^2 = r^2$$

satisfies

$$T \leq \frac{v}{2(1-b/2)} \log \frac{MR^2}{mr^2} \quad (56)$$

As an application of this theorem, an example of a forced oscillation is considered.

**Example—**Consider the system

$$\dot{u} = pu + \varepsilon U(u, \dots, u_n) + F(t) \quad (57)$$

where  $p$  is assumed to have all its eigenvalues with negative real parts,  $\varepsilon$  is a small parameter,  $U$  is continuously differentiable and  $F(t)$  is the forcing term with period  $\tilde{T}$ .

Let the system (56) have a periodic solution

$$u_s = u_s^0(t), \quad u_s^0(t) = u_s^0(t + \tilde{T}) \quad (s = 1, \dots, N) \quad (58)$$

and let the linear transformation

$$y = Cu \quad (59)$$

transform the system  $\dot{u} = pu$  into its normal form

$$\dot{y} = Jy \quad (60)$$

By means of the transformation (59) the system (56) was reduced to a system

$$\dot{y} = Jy + \varepsilon Y(y) + \Phi(t) \quad (61)$$

where  $\Phi(t) = CF(t)$  has the same period as  $F(t)$ . Under this transformation, the periodic solution (58) is reduced to

$$y_s = y_s^0(t) = \sum_{\sigma=1}^N \cos u_\sigma^0(t) \quad (62)$$

Consider the perturbations  $x_s = y_s - y_s^0(t)$  then  $x$  satisfies

$$\dot{x} = Jx + \varepsilon q(t)x + \varepsilon X(x, t) \quad (63)$$

where  $q(t)$  is a periodic matrix with period  $T$ , and it may be evaluated through  $Y(t)$  and  $y_s^0(t)$ . If one takes  $\xi = C^{-1}x$ , then  $\xi = u - u^0(t)$  is the perturbation vector in  $u$  space.

By means of the above method the matrix  $A$  is constructed, with its maximum eigenvalue  $v$ , and the maximum and minimum eigenvalues of the matrix  $C'AC$  are  $M$  and  $m$  respectively. The following results are obtained.

**Theorem 7.** The decaying time  $T$  of the motion of the system (57) from an initial point in the  $R$  neighbourhood of the periodic solution (58) to a point in the  $r$  neighbourhood of the periodic solution (58) satisfies

$$T \leq \frac{\nu}{2 \left[ 1 - \frac{(b+c)}{2} \varepsilon \right]} \log \frac{MR^2}{mr^2} \quad (64)$$

where the term  $X$  in (62) satisfies

$$|\text{grad } \bar{V} \cdot X| < bx'x \quad (b < 2, \bar{V} = x'Ax) \quad (65)$$

and  $C$  is the maximum eigenvalue of the matrix  $q(t) + [q(t)]$  when  $t \in [0, \tilde{T}]$ . By the above-mentioned  $r$  neighbourhood of the periodic solution (58), is meant the set of points which satisfies the inequality

$$\sum_{s=1}^N [u_s - u_s^0(t)]^2 \leq r^2$$

#### On the Estimation of Decaying Time for a Quasi-reducible Linear System

In the study of the dynamic systems, one may sometimes fail to approximate it by a linear system with constant coefficients. In this case one may take a reducible system as its approximations, and construct the corresponding Liapunov function and estimate the decaying time.

Consider the non-linear system

$$\dot{x} = p(t)x + X(x, t) \quad (66)$$

Let the linear approximation system

$$\dot{x} = p(t)x \quad (67)$$

be a reducible system. Assuming the characteristic number to be all positive, there is a Liapunov transformation

$$y = C(t)x \quad (68)$$

which transforms the system (67) into its real normal form

$$\dot{y} = Jy \quad (69)$$

By means of the method mentioned in the previous section,

$$V = x'C'ACx \quad (70)$$

is taken as a Liapunov function for the system (67). Since (68) is the Liapunov transformation when  $t \geq t_0$ , the maximum and minimum eigenvalues  $M$  and  $m$  cannot equal zero. Obviously, the maximum eigenvalue  $\nu$  of  $A$  can be calculated through the characteristic numbers of (67) by the same method. Parallel to Theorem 2 the following results may be obtained.

**Theorem 8.** The decaying time  $T$  of the motion of the system (67) from an initial point in the sphere

$$\sum_{s=1}^N x_s^2 = R^2$$

to a point in the sphere

$$\sum_{s=1}^N x_s^2 = r^2$$

satisfies

$$T \leq \frac{\nu}{2} \log \frac{MR^2}{mr^2} \quad (71)$$

Furthermore, if  $X$  satisfies the condition

$$|\text{grad } V \cdot X| < bx'C'Cx \quad (b < 2) \quad (72)$$

then one has the following results.

**Theorem 9.** The decaying time  $T$  of the motion of the system (66) from an initial point in the sphere

$$\sum_{s=1}^N x_s^2 = R^2$$

to a point in the sphere

$$\sum_{s=1}^N x_s^2 = r^2$$

satisfies

$$T \leq \frac{\nu}{2(1-b/2)} \log \frac{MR^2}{mr^2} \quad (b < 2) \quad (73)$$

where  $X$  satisfies (72).

Since the linear system with periodic coefficients is a reducible system, and its characteristic number can be represented through its characteristic exponentials, the results in this section can be applied to the general periodic systems.

#### Appendix

##### Proof of the Fundamental Theorem

Obviously the system is stable with respect to (2) taking on zeros.

For any given  $\eta > 0$  the region  $H \geq \|\phi\|_{\kappa} \geq \eta$  is considered. From the conditions mentioned in the theorem, the function  $V$  takes on maximum  $M > 0$  and minimum  $m > 0$  and the negative definite function  $dV/dt|_{1.1}$ , with respect to (2), takes on maximum  $-\alpha < 0$ . Let  $T = (M - m)/\alpha + 2\beta$ , where  $\beta$  is any arbitrary positive number. This is the required  $T$  and it is independent of the initial conditions. Thus, part (A) of the fundamental theorem holds.

Parallel to Theorem 3, from the conditions of the fundamental theorem, the following lemmas can be proved.

**Lemma 1.** If the system (1) is equi-asymptotically stable with respect to (2), taking zeros for the initial values in  $\mathcal{F}(n-k)(H)$ , then there is  $\psi(\tau)$  such that the motions, defined by the initial points of the above-mentioned region, satisfy

$$(a) \quad \|\phi[x(t_0 + \tau, x_1^0, \dots, x_n^0, t_0)]\|_{\kappa} \leq \psi(\tau)$$

$$(b) \quad \lim_{\tau \rightarrow \infty} \psi(\tau) = 0 \quad \psi'(\tau) \leq 0 \quad \tau > 0$$

**Lemma 2.** For any given two positive functions  $M(\eta)$  and  $\psi(\eta)$ , where  $M(\eta)$  is an increasing function and  $\lim \psi(\eta) = 0$  there is a function  $G(\eta)$  such that

$$(a) \quad G(\eta) > 0, \quad G'(\eta) > 0 \quad \text{as } \eta > 0$$

$$(b) \quad G(0) = G'(0) = 0$$

$$(c) \quad \int_0^{\infty} G[\psi(\tau)] d\tau < \infty \quad \int_0^{\infty} G'[\psi(\tau)] M(\tau) d\tau < \infty$$

**Lemma 3.** If the system (1) is equi-asymptotically stable with respect to (2) taking on zeros, and the rank of the matrix (19) is  $K$ , then there are two positive constants  $A$  and  $\lambda$  independent of  $t_0$  and  $\phi^0$  such that

$$\left| \frac{\partial \phi^2}{\partial t_0} \right| < A e^{\lambda \tau} \quad \left| \frac{\partial \phi^2}{\partial \phi_s^0} \right| < A e^{\lambda \tau}$$

where

$$\phi^2 = \sum_{s=1}^K \phi_s^2(t, \phi_1^0, \dots, \phi_n^0, t_0)$$

$\phi_i^0$  ( $i = 1, \dots, n$ ) are the initial values and  $t$  is replaced by  $t_0 + \tau$ . Parallel to Theorem 3, the Liapunov function may be taken as

$$V = \int_0^\infty G \left[ \sum_{s=1}^K \phi_s^2(t + \tau, \phi_1, \dots, \phi_n, t) \right] d\tau$$

then one has

$$\left| \frac{\partial V}{\partial \phi_0} \right| = \left| \int_t^\infty G' \left[ \phi^2(t + \tau, \phi_1, \dots, \phi_n, t) \right] \frac{\partial \phi^2}{\partial \phi_i} d\tau \right| < \infty$$

( $i = 1, \dots, K$ )

It is convergent and uniformly bounded. This implies that  $V$  satisfies the Lipschitz condition and thus  $V$  has an infinitely small upper bound.

$$V > \frac{1}{2L} [\phi_1^2 + \dots, \phi_K^2] G \left[ \frac{1}{2} \cdot \sum_{i=1}^K \phi_i^2 \right]$$

where  $L$  is the Lipschitz constant. It implies  $V$  to be positive definite

$$\left. \frac{dV}{dt} \right|_{1,1} = -G[\phi_1^2 + \dots + \phi_K^2]$$

Hence

$$\left. \frac{dV}{dt} \right|_{1,1}$$

is negative definite with respect to (2).

The proof is complete.

## References

- <sup>1</sup> LIAPUNOV, A. M. *The General Problem of Stability of Motion*. 1950. Gostekhizdat
- <sup>2</sup> MASSERA, J. L. Contributions to stability theory. *Ann. Math.* 64, No. 1 (1956)
- <sup>3</sup> MALKIN, I. G. Contribution to the theory of the invertibility of Liapunov's theorem on asymptotic stability. *Prikl. Mat. i Mekh.* XVIII, B<sub>2</sub> (1954)
- <sup>4</sup> ZUBOV, V. I. *The Methods of A. M. Liapunov and their Application*. 1957. Leningrad; Gos. Univ.
- <sup>5</sup> KRASOVSKIY, N. N. *Certain Problems in the Theory of Stability of Motion*. 1959. Fizmatgiz
- <sup>6</sup> CH'ING YÜAN-HSÜN. *A Lecture on the General Problem of Stability of Motion*. 1958. Peking
- <sup>7</sup> KALMAN, R. E. and BERTRAM, J. E. Control system analysis and design via the 'Second Method' of Liapunov. *Trans. Amer. Soc. mech. Engrs Ser. D*, 82, No. 2 (1960)
- <sup>8</sup> RUMYANTSEV, V. V. Stability of motion in relation to part of the variables. *Vestnik. Moskov. Univ.* B<sub>4</sub> (1957)
- <sup>9</sup> HWANG, L. Problems in estimation of decaying time for multi-dimensional non-linear systems. *Acta Sci. Natur. Peking Univ.* VI, No. 1 (1960)
- <sup>10</sup> CHETAYEV, N. G. The choice of parameters for a stable mechanical system. *Prikl. Mat. i Mekh.* 15, B<sub>3</sub> (1951)
- <sup>11</sup> LETOV, A. M. *The Stability of Non-linear Controlled Systems*. 1955. Gostekhizdat
- <sup>12</sup> CHANG SSU-YING. Estimated solutions to systems of differential equations for accumulation, perturbation and stability of motion, over a finite time interval. *Prikl. Mat. i Mekh.* Vol. XXIII, B<sub>4</sub> (1959)
- <sup>13</sup> TROYTSKIY, V. A. Canonical transformations of the equations of automatic control theory. *Prikl. Mat. i Mekh.* XXI, B<sub>4</sub> (1957)
- <sup>14</sup> CHETAYEV, N. G. *Stability of Motion*. Gostekhizdat

## DISCUSSION

J. P. LASALLE, and G. SZEGÖ, *Research Institute for Advanced Studies, Baltimore 12, Maryland, U.S.A.*

The theory developed in the first part of the paper has some aspects which may be of practical importance in the investigation of the stability properties of invariant manifolds. There are, however, some questions concerning the meaning and significance of the stability concepts introduced and the validity of the 'fundamental theorem' when the manifolds are non-compact.

The stability concepts of the paper are relative to the system of neighbourhoods  $\mathcal{F}(H)$  defined by eqn (4). Because of this, these stability concepts can have, for example, the peculiar property that an invariant manifold can be asymptotically stable in the sense of Liapunov and yet not be asymptotically stable in the sense of this paper.

The difficulty is that the neighbourhoods  $\mathcal{F}(H)$  may contain points which are near infinity. The definitions of the paper are then statements not only about motions near the manifold, but also about motions that are far away from the manifold. The most obvious way to avoid this difficulty is to restrict the definitions to a Euclidean neighbourhood of the manifold. The difficulty can also be avoided by assuming that limiting is positive as the distance from the mani-

fold approaches infinity. We assume throughout the remainder of this discussion that one of these restrictions is in force and that the function is continuous.

The next point we wish to make has to do with the 'fundamental theorem'. When the manifold is compact it is not difficult to show that the neighbourhood system of the paper is equivalent to the Euclidean neighbourhood system of the manifold. Thus for compact manifolds the stability concepts of the paper are equivalent to the usual Liapunov stability concepts for invariant manifolds.

In the case of non-compact manifolds there is a difficulty with the 'fundamental theorem'.

All of this is not to say that the 'fundamental theorem' is not of interest, and it may well have practical applications. In this regard we would like to ask if examples have been worked out which illustrate better than the example of the paper the power of the theorem.

It would also be of interest to know how the theorem needs to be modified in order to be valid for non-compact manifolds. Here, too, it is relevant to ask: what is the relation between the stability concepts of the paper and the usual Liapunov stabilities relative to Euclidean distance in the original state space? We wonder if the author has any results in this direction.

L. HWANG, *in reply*

I thank Professor Lasalle for his comments and for pointing out the following:

(1) In this paper stability is defined according to whether the norm  $\|\varphi\|$  approaches zero or not, rather than according to whether the ordinary Euclidean distance to the given manifold approaches zero or not. It is true that the phase point may have its limit point at infinity and thus the corresponding Euclidean distance may be increasing; it is not even difficult to give illustrative examples. However, since in practice we are usually interested in  $\varphi$  rather than the Euclidean distance, the latter concept is not introduced here. In fact, the norm defined in this paper is the distance in a more generalized space.

(2) The stability definition and the method of treatment in this paper is a generalization of Liapunov's definition and his second method. It is possible that a certain system is asymptotically stable in the sense of  $\varphi$ , as defined in this paper, while it is not asymptotically stable in Liapunov's sense. The converse case is also possible. The definition of stability here is not equivalent to that of Liapunov.

(3) The proof of the fundamental theorem does not demand the manifold itself to be compact. This theorem holds valid so long as the prescribed conditions are satisfied. The system given in Professor Lasalle's comments  $\dot{x}_1 = x_1^2$ ,  $\dot{x}_2 = -x_2$  does satisfy the definition of asymptotic stability relative to  $\varphi \equiv x_2 = 0$  since  $x_2 = x_2^0 e^{-t}$ , although the curve in the  $x_1 x_2$  plane may not approach to the  $x_1$  axis in the ordinary geometrical sense (Figure A). Moreover, for practical systems it is usually assumed that the motion does not approach to infinity in a finite time interval.

Finally, I would like very much to know Professor Lasalle's works, if any, on the relation between the stability concept for invariant manifolds and the ordinary Liapunov's stability concept.

B. S. RAZUMICHIN, *Institute of Mechanics, Academy of Sciences, Moscow, U.S.S.R.*

I have two questions to ask the author.

(1) What is the meaning of the statement that all remaining coordinates are free?

(2) What method is used to prove that the variety (complexity) under investigation is not empty?

I fully agree with Dr. Lasalle's remarks. The definition given in the paper differs from Liapunov's general definition only by the fact that the initial values of some quantities have no restrictions. This is obtained at an extremely high price, because it is absolutely necessary to prove that the variety under investigation is not empty, apart from stability investigations.

I cannot agree with the statement that the estimation 45 (Theorem 4) is exact, as the inequality becomes equality only for two integral curves and at one fixed moment. On all other points of time axes the exact inequality is satisfied.

I draw the author's attention to my paper<sup>1</sup> which shows the estimate for separate coordinates obtained by Jacoby's transformation.

#### Reference

- <sup>1</sup> RAZUMICHIN, B. S. *Prikl. Mat. i Mekh.* XXI, No. 1 (1957)

L. HWANG, *in reply*

(1) It is pre-assumed that  $\varphi_i = 0$  ( $i = 1, 2, \dots, K$ ) is non-empty, which corresponds to the standard working state in practical application.

(2) The inequality (45) also holds valid for a linear system with constant coefficients. In fact, when the system is reduced to its Jordan canonical form by means of linear transformation, there must exist two particular motions corresponding to the upper and the lower bounds of this inequality.

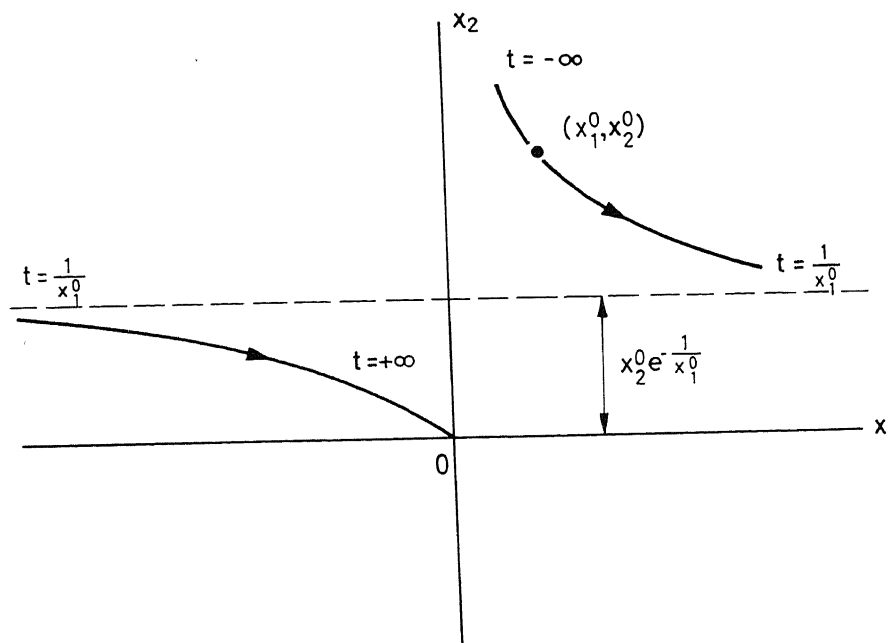


Figure A

# New Methods for Constructing Liapunov Functions for Time-invariant Control Systems

G. P. SZEGÖ

## Summary

In the investigation of the stability properties of non-linear control systems one likes to have a general procedure for constructing Liapunov functions which will always yield a solution of the stability problem. The only general method in existence up to now was proposed by Zubov in 1955 and it is based upon the integration of certain quasi-linear partial differential equations.

In this contribution the author generalizes and extends the work of Zubov, and proposes new types of partial differential equations which, as the Zubov's partial differential equation, will provide a solution of the stability problem. On the other hand, these equations allow a certain freedom of choice of some parameters and some functions. The method developed, here will also give a nice intuitive interpretation of the meaning of Liapunov functions and of some problems of optimal control.

## Sommaire

Dans l'étude des propriétés de stabilité des systèmes de commande non-linéaires il est très utile d'avoir une méthode générale de construction des fonctions de Liapunov qui fournissent toujours une solution au problème de stabilité. A l'heure actuelle la méthode proposée par Zubov (1955) est la seule à remplir ces conditions; cette méthode est basée sur l'intégration de certaines équations quasi-linéaires aux dérivées partielles.

Dans la présente contribution l'auteur généralise et étend les travaux de Zubov et présente de nouveaux types d'équations aux dérivées partielles qui, comme celle de Zubov, résoudre le problème de stabilité. Ces nouvelles équations laissent par ailleurs une certaine liberté quant au choix de certains paramètres et de certaines fonctions. La méthode proposée donne de plus une interprétation intuitive et élégante de la signification des fonctions de Liapunov et de certains problèmes de commande optimale.

## Zusammenfassung

Bei der Untersuchung der Stabilitätseigenschaften nichtlinearer Regelungssysteme möchte man ein allgemeines Verfahren zur Konstruktion Liapunovscher Funktionen haben, welches immer eine Lösung des Stabilitätsproblems liefert. Die einzige bisher existierende allgemeine Methode wurde 1955 von Zubov vorgeschlagen, sie beruht auf der Integration gewisser quasilinearer partieller Differentialgleichungen.

In diesem Beitrag wird die Arbeit von Zubov verallgemeinert und erweitert und neue Typen von partiellen Differentialgleichungen vorgeschlagen, die ebenso wie die von Zubov eine Lösung des Stabilitätsproblems liefern. Andererseits lassen diese Gleichungen eine gewisse Freiheit in der Wahl einiger Parameter und Funktionen. Die hier entwickelte Methode gibt auch eine gute anschauliche Interpretation der Bedeutung der Liapunov-Funktionen und von Problemen der optimalen Regelung.

## Introduction

In this paper some new techniques leading to a new approach to the problem of stability analysis of time-invariant control systems are developed.

This study is limited to the investigation of the stability properties of completely defined control systems. The new method for stability investigation which is the final outcome of this work will, in principle, always yield some solution of the stability problem. The price that must be paid for assuring that the method always works is the restriction to a particular class of Liapunov functions<sup>1, 2</sup>. These Liapunov functions are solutions of a partial differential equation which turns out to be the generalization of an analogous equation proposed by Zubov<sup>3</sup>. There exist obvious connections of this method with the problem of optimal control.

The research was carried out while the author was visiting the Control and Information Systems Laboratory of the School of Electrical Engineering, Purdue University and was supported in part by the National Science Foundation Grant No. G-16460 and in part by the Office of Naval Research Nonr-3693 (00).

## Nomenclature

Standard vector notation is used with the following conventions: capital letters are matrices, small Latin letters are vectors, Greek letters and small Latin letters with subscripts are scalars. Exceptions:  $t, v$ , are scalars.

The inner product of the two vectors  $x$  and  $y$  is denoted by  $(x, y)$ .

The transpose of a matrix  $A$  is denoted by  $A'$ .

An equilibrium state that is stable, but not asymptotically stable, is called weakly stable.

A scalar function  $\theta = \theta(x)$  is positive (negative) definite on the trajectories of a system in a region  $S \subset E_n$  if  $\theta(x) \geq 0$  in  $S$  ( $\theta(x) \leq 0$ ) and  $\theta(x) \neq 0$  on any non-singular solution of the system. By a Liapunov function is meant any scalar function which gives the answer to the stability properties of a solution of a system.

Unless otherwise stated, it is assumed throughout that all the scalar functions used have continuous first partial derivatives.

## Liapunov's Second Method

Liapunov's second method can be codified in a set of theorems<sup>1, 2, 4, 5</sup> which prove that if for a given system there exists a scalar function with certain properties, called a Liapunov function, then conclusions can be drawn about the stability properties of the solutions of the system. There also exist a set of inverse theorems<sup>4, 6</sup> which guarantee the existence of such a function.

Given the non-linear autonomous dynamic system<sup>7</sup>

$$\dot{x} = f(x) \quad f(0) = 0 \quad (1)$$

where  $f(x)$  is a vector-valued function, the problem of the stability analysis of its equilibrium point is then formally



reduced to the search for a positive definite<sup>8</sup> scalar function  $\psi = \psi(x)$  and a scalar function  $v = v(x)$ ,  $v(0) = 0$ , such that the partial differential equation

$$\psi(x) = -(\text{grad } v(x), f(x)) = \sum_{i=1}^n \frac{\partial v}{\partial x_i} f_i(x) \quad (2)$$

is satisfied.

From the form of the scalar function  $v(x)$ , conclusions may be drawn about the stability properties of the solution  $x = 0$  of the system of eqn (1) and about the range of these properties<sup>9, 10</sup>.

The major problem is then to find a definite scalar function  $\psi(x)$ , such that the solution  $v = v(x)$  of eqn (2) satisfies the condition  $v(0) = 0$ . The inverse theorems<sup>4, 6</sup> guarantee that such scalar functions  $\psi(x)$  and  $v(x)$  exist. The stability problem is then reduced to the problem of finding necessary and sufficient conditions that a given scalar function  $\psi(x)$  be positive definite. The existence and the possibility of using such conditions is however much in doubt.

A different and more sensible approach is that of finding a sufficient condition which guarantees that a scalar function  $\psi(x)$  is at least definite on the trajectories of system (1) and such that there always exists a function  $v(x)$  satisfying this condition and a corresponding  $v = v(x)$ ,  $v(0) = 0$  which satisfies eqn (2).

#### A Generalization of Zubov's Equation

In this section some remarks are made on the form of the scalar function  $\psi(x)$ , defined in eqn (2). These remarks will constitute the basis of the present method.

The following question is posed. If an arbitrary scalar function  $v_1 = v_1(x)$  is given and

$$\psi_1(x) = (\text{grad } v_1(x), f(x)) \quad (3)$$

under what conditions on  $v_1(x)$  is it possible to compute from the scalar function  $v_1 = v_1(x)$  a new scalar function  $v_2 = v_2(x)$ , such that  $\psi_2(x) = (\text{grad } v_2(x), f(x))$  has a certain required form?

Consider the Pfaffian differential equation

$$\omega(x) = (y(x), dx) = 0 \quad (4)$$

Let  $\mu(x)$  be an integrating factor of this equation, that is

$$\mu(x) y(x) = \text{grad } v(x) \quad (5)$$

The scalar function  $v = v(x)$  is then a particular integral of the Pfaffian differential eqn (4), that is,

$$dv = (\text{grad } v(x), dx) \quad (6)$$

Consider now an arbitrary function  $\alpha = \alpha(v)$ . From eqn (4) it follows that

$$\mu(x) \frac{d\alpha}{dv} (y(x), dx) = 0 \quad (7)$$

By substituting eqn (5) into eqn (7)

$$\frac{d\alpha}{dv} (\text{grad } v(x), dx) = 0 \quad (8)$$

which from eqn (6) is identically equal to

$$\frac{d\alpha}{dv} dv = d\alpha = 0 \quad (9)$$

These results can be summarized in the following.

**Theorem 1**—If  $\mu(x)$  is an integrating factor of the Pfaffian differential eqn (4) with solution  $v = v(x)$  and  $\alpha = \alpha(v)$  is an arbitrary scalar function, then  $\mu(x) d\alpha/dv$  is also an integrating factor of eqn (4) with solution  $\alpha = \alpha(v(x)) = \alpha^*(x)$ .

It has been shown that the stability problem is reduced to the search for scalar functions  $v = v(x)$ ,  $v(0) = 0$  and  $\psi = \psi(x)$  positive definite on the trajectories of system (1). Then the problem may be reduced to that of seeking a scalar function  $\alpha(v)$  such that

$$\frac{d\alpha}{dv} \mu(x) \omega(x) = \psi(x) \quad (10)$$

or

$$\frac{d\alpha(v)}{dv(x)} = \frac{\psi(x)}{\mu(x) \omega(x)} \quad (11)$$

The functional eqn (11) can be readily solved if

$$\frac{\psi(x)}{\mu(x) \omega(x)} = \beta(v(x)) \quad (12)$$

On the basis of these considerations the following procedure can be developed.

Take a scalar function  $v_1 = v_1(x)$ ,  $v_1(0) = 0$  and compute its total time derivative

$$\frac{dv_1}{dt} = (\text{grad } v_1(x), f(x)) = \gamma(x), \gamma(0) = 0 \quad (13)$$

Next look for a scalar function  $\psi(x)$  which is at least definite on the trajectories of system (1) and a scalar function  $\beta(v_1)$ ,  $\int_0^{v_1} \beta(s) ds < \infty$  such that

$$\frac{\psi(x)}{\gamma(x)} = \beta(v_1) \quad (14)$$

Then the differential equation

$$\frac{d\alpha(v_1)}{dv_1} = \frac{\psi(x)}{\gamma(x)} = \beta(v_1) \quad (15)$$

can be integrated. Its solution  $\alpha = \alpha(v_1) = \alpha^*(x)$  will be such that

$$\dot{\alpha}^* = (\text{grad } \alpha^*(x), f(x)) = \psi(x) \quad (16)$$

which is at least definite on the trajectories of the system (1), and because of the assumptions made on  $\alpha(v_1)$  and  $v_1(x)$  solve the stability problem.

Substituting eqn (14) into eqn (13) gives

$$\frac{dv_1}{dt} = (\text{grad } v_1(x), f(x)) = \frac{\psi(x)}{\beta(v_1)} \quad (17)$$

which is a generalization of Zubov's eqn<sup>8</sup>:

$$(\text{grad } v, f(x)) = \phi(x)(1+v) \quad (18)$$

The stability theorem deduced from eqn (18) may be stated as:

**Theorem 2**—The stability problem of the solution  $x = 0$  of the system eqn (1) is reduced to finding scalar functions  $v_1(x)$ ,  $\psi(x)$ ,  $\beta(v_1)$  such that  $v_1(0) = 0$ ,  $\int_0^{v_1} \beta(s) ds < \infty$  and  $\psi(x)$  is definite on the trajectories of the system of eqn (1).

Integrating eqn (15)

$$\alpha(v_1) = \int_0^{v_1} \beta(s) ds \quad (19)$$

from which is deduced

**Theorem 3**—The solution  $x=0$  of the system of eqn (1) is asymptotically stable in a closed, bounded region  $S: \alpha^*(x) \leq \delta$ ,  $\delta > 0$ , if there exist scalar functions  $v_1(x)$ ,  $\psi(x)$ ,  $\beta(v_1)$  satisfying the following conditions:

- (i)  $v_1(0) = 0$ .
  - (ii)  $\psi(x)$  negative definite on the trajectories of the system eqn (1).
  - (iii)  $\int_0^{v_1} \beta(s) ds < \infty$ .
  - (iv)  $\alpha^*(x) = \int_0^{v_1} \beta(s) ds > 0$  in  $S$ ,  $x \neq 0$ ,  $\alpha^*(0) = 0$ ,
- and such that eqn (18) is satisfied.

**Corollary 1**—The solution  $x=0$  of the system of eqn (1) is asymptotically stable in the large if all the conditions of Theorem 2 are satisfied and

$$\lim_{\|x\| \rightarrow \infty} \alpha^*(x) = \lim_{\|x\| \rightarrow \infty} \int_0^{v_1} \beta(s) ds = \infty$$

**Corollary 2**—If all the conditions of Theorem 3 are satisfied with the sign of  $\psi(x)$  changed, then the solution  $x=0$  of the system of eqn (1) is completely unstable<sup>11</sup>.

**Remark 1**—It is always possible to give sufficient conditions for complete instability, from any theorem on asymptotic stability. In the following sections are presented numerous theorems on asymptotical stability, and it is always implied, even if not explicitly stated, that a similar theorem for complete instability holds.

Since, given any scalar function  $v_2 = v_2(x)$ , it is always possible to find a functional  $\Omega = \Omega(v_2(x))$ ,  $\Omega(0) = 0$  such that the scalar function  $\Omega = \Omega^*(x) = \Omega(v_2(x))$  is semi-definite, the following simplified procedure for constructing Liapunov's functions can be developed.

First, seek a scalar function  $v_2 = v_2(x)$ ,  $v(0) = 0$ , such that

$$\frac{dv_2}{dt} = (\text{grad } v_2(x), f(x)) = \theta(v_2) \quad (20)$$

where  $\theta = \theta(v_2)$ , is a bounded scalar function. Eqn (20) is a special case of eqn (17).

Then it is always possible to integrate the equation

$$\frac{d\alpha_2(v_2)}{dv_2} = \frac{\Omega(v_2)}{\theta(v_2)} = \beta_2(v_2) \quad (21)$$

and its solution  $\alpha_2 = \alpha_2(v_2) = \alpha_2^*(x)$  will be such that

$$\frac{d\alpha_2}{dt} = (\text{grad } \alpha_2^*(x), f(x)) = \Omega^*(x)$$

where  $\Omega^*(x)$  is semidefinite. If no degeneracy occurs and if

$$\int_0^{v_2} \beta_2(x) < \infty$$

then the scalar function  $\alpha_2 = \alpha_2^*(x)$  is a Liapunov function of the system of eqn (1). Theorems 1 and 2 and the corollaries apply with minor changes to this case.

The following theorem answers the question of the existence of solutions of eqn (20).

**Theorem 4**—There always exists a scalar function  $\theta(v_2)$  such that eqn (20) has a solution which satisfies the condition

$v_2(0) = 0$ . In particular if the system of eqn (1) is asymptotically stable  $v_2(x)$  is definite and  $\theta(v_2)$  may be chosen so that  $\theta(v_2) = \lambda v$ ,  $\text{Re}\{\lambda\} < 0$ .

**Proof**—This theorem may be proved directly following a method used by Vrkoch<sup>12</sup>, or by using the inverse theorem<sup>4, 6</sup>. A proof is sketched using this latter way for the case of asymptotic stability.

Given any Liapunov function  $v = v(x)$ ,  $v(0) = 0$ ,  $\dot{v}(x) \neq 0$  for  $x \neq 0$ ,  $v(0) = 0$ , this function represents a hypersurface in  $E_{n+1}$  with a strong minimum for  $x = 0$ .

Take any section of this hypersurface with the hyperplane  $v = \text{const} \neq 0$ . This section is a closed bounded hypersurface. Represented in parametric form

$$x_i = x_i(t_1, \dots, t_{n-1}) \quad (22)$$

Since  $\dot{v}(x) > 0$  for  $x \neq 0$  it is possible to construct a unique integral surface  $S = v^*(x)$  of the equation

$$\dot{v}^* = (\text{grad } v^*(x), f(x)) = -v^*(x) \quad (23)$$

going through the hypersurface (eqn (22)).

Consider the characteristic system of the partial differential eqn (23)

$$\frac{dv^*}{-v^*} = dt, \quad \frac{dx_i}{f_i(x)} = dt \quad i = 1, \dots, n \quad (24)$$

Since the system of eqn (1) is by assumption asymptotically stable, the solutions  $x_i = x_i(t)$ ,  $v^* = v^*(t)$  of the characteristic system tends to zero as  $t \rightarrow \infty$ . Hence  $v^*(x)|_{x=0} = 0$ .

**Remark 2**—An existence theorem analogous to Theorem 4 may be proved also for the partial differential eqn (17). It is seen that the necessary and sufficient conditions that the scalar function  $\psi(x)$  must satisfy for eqn (17) to have a solution  $v_1 = v_1(x)$  such that  $v_1(0) = 0$  are not identical to the conditions on  $\theta(x)$  in Zubov's eqn (18). In this latter case  $\int_0^t \phi(x(\tau)) d\tau < \infty$ , a relation which in this case is only sufficient.

It is worthwhile to emphasize the major differences between eqns (17) and (20). As previously pointed out the  $v$ -functions  $\alpha = \alpha^*(x)$ , obtained from eqn (18) are Liapunov functions for the system of eqn (1). The only possible degenerate case, in which  $\alpha$  is semi-definite, will never arise.

In fact if  $\alpha = \alpha^*(x)$  is continuous and semi-definite then it has a strong minimum on the manifold  $M$  on which  $\alpha^*(x) = 0$ . Hence  $\text{grad } \alpha^*(x) = 0$  on  $M$  and  $M$  is an integral manifold of system (1)<sup>3, 9</sup>. This contradicts the hypothesis that  $\psi(x)$  is definite on the trajectories of system (1). This reasoning does not apply to eqn (20) where  $\Omega^*(x)$  is only semi-definite. Then  $\alpha^*(x) = \Omega^*(x) = 0$  on some integral manifold  $N$ . All the information obtained in this case concerns the stability of the manifold  $N$  and not of the equilibrium point. In some cases then the procedure must be repeated in order to find a new  $\alpha_1 = \alpha_1^*(x)$  satisfying an equation of the type of eqn (17) which may be semi-definite as long as there does not exist a point  $x = x_e \neq 0$  on which  $\alpha^*(x_e) = \alpha_1^*(x_e) = 0$ . In this case one not only obtains information about the stability properties of the equilibrium point, but one is also able to find some integral of the system. The same situation may arise from eqn (17) if  $\psi(x)$  is a semi-definite function.

Eqn (20) is very important in itself since its solutions are the so-called isochrones of system (1). The knowledge of the isochrones gives important information about the qualitative behaviour of the solutions of the system of eqn (1).

**Remark 3**—For some systems it may happen that the scalar function  $\psi(x)$  in eqn (17) or  $\Omega^*(x)$  in eqn (20) is identically zero. The scalar function  $\alpha = \alpha^*(x)$  is then a first integral of the system.

### A Useful Change of Variables

The method for constructing Liapunov functions developed in the previous section contrasts strongly with the methods in use up to now. The essence of this method is the introduction of the functions  $\beta(v_1)$  and  $\theta(v_2)$  respectively in eqns (17) and (20). The major step is now to find a scalar function  $v = v(x)$ ,  $v(0) = 0$ , such that  $dv/dt$  has the form  $\theta_2(v)$  or  $\psi(x)\beta(v)$ , but is otherwise completely arbitrary.

Regarding the problem of constructing a Liapunov function by solving a partial differential equation it is seen that the original linear partial differential eqn (2) (with unknown right-hand side) has become a quasi-linear partial differential eqn (17) or (20) whose right-hand side has a certain well-defined form.

In this paragraph is performed a particular transformation of variables  $x \rightarrow z$ . One of the components of the next state vector  $z$  is the scalar function  $v$ .

The stability problem will then be reduced to a search for a scalar function  $\xi = \xi(z)$ , satisfying a certain non-linear partial differential equation, whose right-hand side is any definite function which depends on only one particular component of the vector  $z$ .

Consider the scalar equation

$$w = v(x) \quad v(0) = 0$$

which is equivalent to the equation

$$x_i = \xi_i(x_1, \dots, x_{i-1}, x_{i+1}, \dots, x_n, w) \quad (25)$$

Since this operation is performed only once in the whole procedure the notations can be simplified. Introduce the  $n$  vector  $z$  defined as follows

$$z_k = x_k \quad k \neq i \quad z_i = w \quad (26)$$

Eqn (32) may be written

$$x_i = \xi_i(z) \quad (27)$$

so that  $v(x_1, \dots, x_{i-1}, \xi_i(z), x_{i+1}, \dots, x_n) = w = 0$  from which

$$\frac{\partial v}{\partial x_j} + \frac{\partial v}{\partial x_i} \frac{\partial \xi_i}{\partial x_j} = 0 \quad (i \neq j) \quad (28)$$

$$\frac{\partial v}{\partial x_i} \frac{\partial \xi_i}{\partial w} = 1$$

and finally

$$\frac{\partial v}{\partial x_i} = 1 \left/ \frac{\partial \xi_i}{\partial w} \right. \quad (29)$$

$$\frac{\partial v}{\partial x_j} = - \frac{\partial \xi_i}{\partial z_j} \left/ \frac{\partial \xi_i}{\partial w} \right. \quad (i \neq j)$$

The transformation of coordinates

$$x \rightarrow z, \xi_i(z) \quad (30)$$

can now be performed on eqns (17) and (20).

These equations take respectively the form

$$\left( f_i(x) - \sum_j \frac{\partial \xi_i}{\partial z_j} f_j(x) \right) \Big|_{x_i = \xi_i} = \frac{\psi(x)}{\beta(x_1)} \Big|_{x_i = \xi_i} \cdot \frac{\partial \xi_i}{\partial w_1} \quad (i \neq j) \quad (31)$$

and

$$\left( f_i(x) - \sum_j \frac{\partial \xi_i}{\partial z_j} f_j(x) \right) \Big|_{x_i = \xi_i} = \theta(w_2) \frac{\partial \xi_i}{\partial w_2} \quad (i \neq j) \quad (32)$$

Eqn (32) is of particular interest and can be written as

$$\left( f_i(x) - \sum_j \frac{\partial \xi_i}{\partial z_j} f_j(x) \right) \Big|_{x_i = \xi_i} \frac{1}{\frac{\partial \xi_i}{\partial w_2}} = \theta(w_2) \quad (j \neq 1) \quad (33)$$

In this latter equation the usefulness of the present method is much emphasized.

It can be seen that the only requirement is that the right-hand side of eqn (33) depends only on  $w_2$ . If the function  $\beta(w_2)$ , defined by eqn (21) satisfies the condition  $\int_0^\infty \beta(s) ds < \infty$ , the problem of the stability of the solution  $x = 0$  of the system under investigation is solved. If one cannot find a scalar function  $\alpha(w_2)$  such that  $\alpha(w_2)/\theta(w_2)$  is bounded, then it is still possible to study the stability of some first integral of the system, going through the origin.

In other words, whatever the function  $\theta(w_2)$  is it will always be possible to have some answer about the stability properties of the system.

Unfortunately the solution of the partial differential equation is not a simple matter and it is possible to integrate it explicitly only if it is possible to integrate its characteristic system. A more reasonable approach is to choose a suitable form for the unknown scalar function  $\xi_i = \xi_i(x)$ ,  $\xi_i(0) = 0$  having a certain number of known coefficients, then compute the unknown coefficients in such a way that the right-hand side of eqn (33) is a function which depends only on  $w_2$ .

The possibility of doing this depends of course on the right choice of the form  $\xi_i$ . Although numerous examples have been solved, no general information about suitable forms for  $\xi_i$  is available.

### Examples

In the case of non-linear systems, if one is not able to integrate the characteristic system of eqns (31) or (32), then an alternative procedure is to look for a suitable form of the unknowns  $v = v(x)$  or  $\xi_i(z, w)$  which allows a separation of variables.

Consider for example the system

$$\begin{aligned} \dot{x} &= y \\ \dot{y} &= -ay - ax^3 - x^2y \end{aligned} \quad (34)$$

for which

$$v_1 = ax + y \quad (35)$$

and

$$\dot{v}_1 = -x^2v_1 \quad (36)$$

It is seen that the solution  $y = -ax$  is asymptotically stable. By integrating it one obtains  $x = x_0 \exp(-at)$ .

It is concluded that the equilibrium point  $x = y = 0$  is asymptotically stable for  $a > 0$  and unstable for  $a < 0$ . In this particular case the stability problem is solved directly from eqns (35) and (36). This is not always the case.

Consider for example

$$\begin{aligned}\dot{x} &= y \\ \dot{y} &= ax + ax^2y - y^3 - y \quad (a > 0)\end{aligned}\quad (37)$$

Assume  $v = ax^2 - y^2$  then  $\dot{v} = 2y^2(1 - v)$

from which no deduction can be drawn. Choose

$$\psi(x) = 2y^2(1 - v)^2$$

then

$$\alpha(v) = \int_0^v \beta(s) ds = \int_0^v (1 - s) ds = v - \frac{1}{2}v^2$$

$$\alpha^*(x) = (ax^2 - y^2) - \frac{1}{2}(ax^2 - y^2)^2$$

$$\dot{\alpha}^*(x) = 2y^2(1 - v)^2$$

from which it is concluded that  $x = y = 0$  is unstable.

The next example illustrates the advantage of eqn (33) with respect to the other formulations. Consider the system

$$\begin{aligned}\dot{x} &= y^3 - x \\ \dot{y} &= x - \frac{1}{2}y\end{aligned}\quad (38)$$

for which eqn (33) takes the form

$$\left[ y^3 - x - \frac{\partial \xi}{\partial y} \left( x - \frac{1}{2}y \right) \right] \frac{1}{\partial \xi / \partial w} = \theta(w) \quad (39)$$

Assume

$$\xi = a[w + f(y)]^{\frac{1}{2}} \quad (40)$$

for which eqn (39) becomes

$$\begin{aligned}\frac{2}{a}y^3[w + f(y)]^{\frac{1}{2}} - 2w - 2f(y) \\ - a \frac{\partial f(y)}{\partial y} [w + f(y)]^{\frac{1}{2}} + \frac{1}{2}y \frac{\partial f(y)}{\partial y} = \theta(w)\end{aligned}$$

from which may be set

$$\theta(w) = -2w \quad (41)$$

$$-2f(y) + \frac{1}{2}y \frac{\partial f}{\partial y} = 0 \quad (42)$$

$$\frac{2}{a}y^3 = a \frac{\partial f}{\partial y} = 0 \quad (43)$$

Integrating eqn (42) gives

$$f(y) = \pm y^4 \quad (44)$$

and from eqn (43)  $f(y) = y^4$  and  $a^2 = \frac{1}{2}$ . Substituting these results into eqn (41) gives

$$\xi = \frac{1}{\sqrt{2}}(w + y^4)^{\frac{1}{2}} \quad (45)$$

which may also be written in the usual form

$$v = 2x^2 - y^4$$

and it may be checked that

$$\dot{v} = \theta(w) = -2v$$

It is concluded that the solution  $2x^2 = y^4$  is globally asymptotically stable.

By direct analysis of this particular solution one can show that the solution  $x = y = 0$  of the system eqn (38) is not globally asymptotically stable.

## Conclusion

The general problem of stability analysis of the equilibrium point of a control system, represented in the form of eqn (1) has been solved.

This general problem has been reduced to the integration of partial differential eqns (20), (31), (17) or (33).

The close-form integration of eqns (20) or (31) presents, in the case of non-linear systems, the same difficulties as the equations which arise from optimal control problems; hence in most of the cases numerical methods must be applied. On the other hand, for expressions of the form of eqns (17) and (33) various examples with closed-form solutions have been found.

## References

- LIAPUNOV, A. Problème général de la stabilité du mouvement. *Ann. Fac. Sci. Toulouse* 2 (1907) 203; reprinted in *Ann. Math., Princeton* Vol. 17
- LA SALLE, J. P., and LEFSCHETZ, S. *Stability of Liapunov's Direct Method*. 1961. New York
- ZUBOV, V. I. Problems in the theory of Liapunov's second method. *Appl. Math. Mech., Leningr.* 19 (1955) 179; *Methods of A. M. Liapunov and their Application*. 1957. Leningrad
- KRASOVSKII, N. N. *Some Problems in the Theory of Stability of a Motion*. 1959. Moscow
- ANTOSIEWICZ, H. A. Survey of Liapunov's second method. *Ann. Math. Stud. Princeton* No. 41 (1958)
- MASSERA, J. L. Contributions to stability theory. *Ann. Math., Princeton* 64 (1956) 182
- KALMAN, R. E. and BERTRAM, J. E. Control system analysis and design via the second method of Liapunov. *J. Basic Engng, Trans. Amer. Soc. mech. Engrs* 82 (1960) 371
- GANTMACHER, F. R. *Theory of Matrices*. 1959. New York; Chelsea
- LA SALLE, J. P. Some extensions of Liapunov's second method. *Trans. Inst. Radio Engrs PGCT-7* (1960) 520
- BARBASHIN, E. A. and KRASOVSKII, N. N. On the existence of a Liapunov function in the case of global asymptotic stability. *Appl. Math. Mech., Leningr.* 18 (1954) 345
- HAHN, W. *Theorie und Anwendung der direkten Methode von Liapunov*. 1959. Berlin; Springer Verlag
- VRKOCH, J. On the inverse theorem of Chetaev. *Czech. Math. J.* 5 (1955) 451

## DISCUSSION

M. MANSOUR, *Institut für Automatik und Industrielle Elektronik ETH, Altbachstrasse 531, Dietlikon ZH, Switzerland*

(1) For the stability study of the system

$$\dot{x} = f(x) \quad (1)$$

$f(0) = 0$ , the main problem is to find  $\psi(x) \partial V(x)$  such that  $\psi(x)$  is definite  $\partial V(x)$  with  $V(0) = 0$  satisfying the basic partial differential equation

$$\psi(x) = \nabla V \cdot f(x) = \sum_i \frac{\partial V}{\partial x_i} f_i(x) \quad i = 1 \dots n \quad (2)$$

Eqn (17) in the paper can be obtained directly from the above equation by assuming

$$V = \int_0^v \beta(s) ds$$

which can be always satisfied by  $V, v$ . Eqn (20) can be obtained directly from (17) by assuming  $\psi(x) = \text{function of } v = \theta(v) \beta(v)$ .

(2) By using eqn (17) the problem of finding a Liapunov function for the study of the equilibrium of the origin  $x = 0$  is not solved as the difficulty of finding two functions  $V(x)$  and  $\psi(x)$  satisfying the basic partial differential equation is not transferred to finding three functions  $v(x)$ ,  $\psi(x)$  and  $\beta(v)$  which I think is still difficult.

(3) As to the example given by the author, the stability or instability of the origin may be detected easily by considering the first approximation. As the right-hand side of the system equations in the three examples is composed only of terms of the first and third degree it is possible to obtain a Liapunov function by solving eqn (2) assuming  $\psi(x) = \psi_2(x) + \psi_4(x)$  and  $V(x) = V_2(x) + V_4(x)$  where  $\psi_2, V_2$  contain only terms of the second degree and  $\psi_4, V_4$  contain only terms of the fourth degree.

(4) Using the Zubov method one can, under certain conditions, find the region of asymptotic stability using his construction procedure either analytically or with the aid of a computer. I do not think that it is possible, using the equations derived by the author, to obtain such a region. However, this method may give information about the stability of some particular solution of the system equations.

G. P. SZEGÖ, *in reply*

I am very grateful to Mr. Mansour for his interest in my work. The disagreement between us is mostly due to misinterpretation and, maybe, incomplete statements of the purpose of this paper. I shall clarify this purpose and present a new theorem that was only mentioned in the paper, and which I hope will emphasize the most important points of my contribution.

The main purpose and the main result of my work is an extension of the class of stability functions beyond the realm of the classical theory. This new class of functions will obviously apply to the investigation of the stability properties of invariant sets of (1) which are larger than equilibrium points. These results will be helpful in the case in which an extremely fine problem of synthesis of sub-optimal control systems exists.

I present the main theorem as follows<sup>1</sup>.

## Theorem

Consider the dynamical system (1). Let

- (i)  $v(x)$  be a scalar function  $\in C^1$  in the whole space  $E^n$ ,
- (ii)  $\psi(v)$  be a scalar function  $\in C^0$  in the whole space  $E^n$ ,
- (iii)  $M$  be the manifold on which  $v(x) = 0$ .

Assume that

- (iv)  $\psi(v(x)) \equiv 0$  in all points of  $M$ ,  $\psi(v(x)) \neq 0$  for  $x \notin M$ ,

- (v) The equation

$$\langle \text{grad } v(x), f(x) \rangle = -\psi(v) \quad (1)$$

is satisfied in the whole space  $E^n$ .

$$(vi) \quad v(x) \psi(v(x)) \geq 0 \quad (2)$$

in the whole space  $E^n$ .

- (vii) The trivial solution  $v = 0$  of the equation

$$\dot{v} = -\psi(v) \quad (3)$$

is globally asymptotically stable.

$$(viii) \quad a(\rho(x, M)) \leq |v(x)| \leq b(\rho(x, M)) \quad (4)$$

where  $\rho(x, M)$  is the Euclidean distance of a point  $x$  from the set  $M$ , the scalar functions  $a(r)$  and  $b(r)$  are positive definite and  $a(r)$  is such that  $\lim_{r \rightarrow \infty} a(r) = \infty$ . Then  $M$  is globally asymptotically stable.

Weaker versions of this problem may be proved for the case of stability and quasi-asymptotic stability; in particular, conditions (3) and (4) will be relaxed. Condition (4), which imposes the complete identity of the  $v$  norm with the Euclidean norm is required for the case that  $M$  is non-compact which is indeed the case of interest. However, when additional informations of the flow may be obtained this condition can be relaxed from either side (as in the examples of the paper).

I hope that this theorem will clarify and strengthen the results of the paper.

As to Mr. Mansour's second and third remarks, it must be clear by now that they are obviously formally correct for the case of stability of equilibrium points, but they do not apply to the general problem studied in the paper.

I do not agree with the fourth remark since I have shown that Zubov's equation is a particular case of the equations derived in the paper.

The first remark is formally correct. I do, however, believe that the derivation of eqns (12) and (20) as done in the paper is worth the trouble because it shows the reason for the assumptions of theorems (2) and (3).

## Reference

- <sup>1</sup> SZEGÖ, G. P. and GEISS, G. R. A remark on a new partial differential equation for the stability analysis of time invariant control systems. *SIAM Journal on Control* 1, No. 4 (1964)

# A Method of Investigating Stability

H. H. ROSENBROCK

## Summary

A method is given for considering the stability of the trivial solution of

$$\dot{x} = A(t, x)x.$$

It is shown that if  $x$  is directed towards the interior of a certain region  $H$  (bounded by hyperplanes) at each of its vertices, whenever  $A$  is within some closed region  $G$  (of the space  $E_n^2$  of the elements of  $A$ ) then, subject to  $A(t, x) \in G$  for all  $t \geq t_0$  and all  $x$ , the solution  $x = 0$  is generally asymptotically and uniformly stable. Some applications are considered.

## Sommaire

On donne une méthode d'étude de la stabilité de la solution du type d'équation:

$$\dot{x} = A(t, x)x.$$

On montre que si  $x$  est dirigé vers l'intérieur d'une certaine région  $H$  (bornée par des hyperplans) à chacun de ses sommets, chaque fois que  $A$  est dans une région fermée  $G$  (de l'espace  $E_n^2$  des éléments de  $A$ ), à condition que  $A(t, x) \in G$  pour tout  $t \geq t_0$  et tout  $x$ , la solution  $x = 0$  est généralement asymptotiquement et uniformément stable. Quelques applications vont être examinées.

## Zusammenfassung

Der Aufsatz behandelt ein Verfahren zur Untersuchung der Stabilität der trivialen Lösung von

$$\dot{x} = A(t, x)x.$$

Es wird gezeigt: wenn  $x$  in einem bestimmten (durch Hyperebenen begrenzten) Bereich  $H$  an jeder seiner Ecken nach innen gerichtet ist und wenn  $A$  sich innerhalb eines abgeschlossenen Bereiches  $G$  (des Raumes  $E_n^2$  der Elemente von  $A$ ) befindet, dann ist unter der Bedingung  $A(t, x) \in G$  für alle  $t \geq t_0$  und alle  $x$  die Lösung  $x = 0$  allgemein asymptotisch und gleichmäßig stabil. Einige Anwendungen werden besprochen.

## Introduction

In an earlier paper by the author<sup>1</sup> some results were given concerning the stability of a second-order differential equation. The method of proof used there can readily be extended in the following way.

The following equation is given:

$$\dot{x} = A(t, x)x \quad (1)$$

where  $x$  is an  $n$ -vector and  $A$  satisfies conditions ensuring that a unique solution passes at time  $t_0$  through any  $x_0$  in some region  $R$  of the space  $(x)$ . Let  $G$  be a closed bounded region in the space  $E_n^2$  of the elements of  $A$ . Let  $H$  be a closed region in  $R$ , such that every point of its boundary lies in at least one of a given set of hyperplanes and such that every ray from the origin intersects the boundary of  $H$  once and only once. (The hyperplanes will be  $n + 1$  or more in number.) Let the points  $x = h_j$  be the vertices of  $H$ , and at each such point let  $\dot{x}$  be non-zero and directed out of the region exterior to  $H$  and into the interior

of  $H$  for all  $A \in G$ . Then every solution of eqn (1), starting at time  $t_0$  from some  $x_0 \in H$ , is uniformly, asymptotically stable subject to the condition that  $A(t, x) \in G$  for all  $t \geq t_0$  and all  $x \in R$ . If  $R$  is the whole space  $(x)$ , the uniform asymptotic stability holds in the large.

To prove this, let  $n$  be the unit inward normal on one face of the region  $H$ . Then  $n' \dot{x}$  is positive everywhere on this face when  $A \in G$ . For suppose to the contrary that at some point  $x$  in the face  $n' \dot{x} \leq 0$  for some  $A_0 \in G$ . The vector  $x$  can be written

$$x = \sum_j \alpha_j h_j \quad (2)$$

where the sum is taken over all the vertices lying in the face considered, every  $\alpha_j \geq 0$ , and some  $\alpha_j > 0$ .

$$\sum_j \alpha_j n' A_0 h_j \leq 0 \quad (3)$$

contradicting the assumption that every  $n' \dot{x}$  is positive at the points  $h_j$  for all  $A \in G$ .

Let  $r = 1$  be some linear dimension of  $H$ , and consider the set of similar regions  $H(r)$ ,  $r \leq 1$ . At corresponding points on the boundary of each such region  $\dot{x}/r$  and  $n$  have the same value for given  $A_0 \in G$ , and  $n' \dot{x}/r$  is positive. For given  $A_0$  and  $r$ , and considering all points lying in the boundary of  $H(r)$ , there is a least value  $\varepsilon(A_0) > 0$  of  $n' \dot{x}/r$ , which is achieved at one of the vertices  $h_j(r)$ . The continuous function  $\varepsilon(A_0)$  is positive for all  $A_0 \in G$ , and hence has a least value  $\varepsilon > 0$  at some point in  $G$ . Thus at every point on the boundary of  $H(r)$ ,  $n' \dot{x} \geq \varepsilon r > 0$  for all  $A \in G$ .

Now let  $V(x)$  be a function defined so that  $V(x) = r$  when  $x$  lies in the boundary of  $H(r)$ . Evidently  $V(x) > 0$  when  $x \neq 0$ , and  $V(0) = 0$ . The function  $V(x)$  has a one-sided derivative in every direction at the boundary of  $H[V(x)]$ , and this derivative is negative along any direction entering  $H$ . Hence  $V$  has a right-hand derivative  $\dot{V}$  (possibly discontinuous) along any solution of eqn (1) and the condition  $n' \dot{x} \geq \varepsilon r > 0$  implies  $\dot{V} \leq -\varepsilon V < 0$ . Then since  $r = V(x)$  we have, for all  $A \in G$ ,  $\dot{V} \leq -\varepsilon V$ . It follows that  $V$  is a Liapunov function for eqn (1) subject to  $A \in G$ , and all solutions starting in  $H(1)$  tend asymptotically and uniformly to  $x = 0$  as  $t \rightarrow \infty$ . The region  $H(1)$  is then to be chosen as large as possible while remaining in  $R$ .

## Examples

Let the vertices of  $H(1)$  be at unit distance from the origin along each (positive and negative) coordinate axis. When  $x_j = 1$  and  $x_i = 0$ ,  $i \neq j$ ,

$$\dot{x}_i = a_{ij}, \quad i = 1, 2, \dots, n$$

If

$$a_{jj} + \sum_{i \neq j} |a_{ij}| \leq -\varepsilon_j < 0 \quad (4)$$

it is easy to see that  $\dot{x}$  is directed towards the interior of  $H$ . If (4) holds for all  $j$ , and if

$$|a_{jj}| \leq \beta, \quad j=1, 2, \dots, n \quad (5)$$

then  $A$  is restricted to a certain closed bounded region,  $G$ , in the space  $E_{n^2}$ . For all  $A \in G$  the conditions of the previous section are fulfilled. Hence if conditions (4) and (5) are satisfied for all  $t \geq t_0$  and all  $x$ , the trivial solution  $x = 0$  is asymptotically and uniformly stable in the large. A Liapunov function for the system is

$$V = \sum_i |x_i| \quad (6)$$

This result may be compared with similar results obtained elsewhere<sup>2,3</sup> by a different method. The earlier results were complicated by the need to allow some of the  $\varepsilon_j$  in (4) to be zero. When this happens the present method allows only stability, and not asymptotic stability, to be asserted. The equations of many physical systems involving the diffusion of heat, matter, electricity, etc., obey the relaxed form of condition (4) used in previous papers by the author<sup>2,3</sup>.

As a second example, let the vertices of  $H(1)$  be at the  $2^n$  points

$$x_1 = \pm 1, x_2 = \pm 1, \dots, x_n = \pm 1 \quad (7)$$

so that at a vertex

$$\dot{x}_i = \sum_j a_{ij} x_j = a_{ii} \operatorname{sgn} x_i + \sum_{j \neq i} a_{ij} \operatorname{sgn} x_j \quad (8)$$

If

$$a_{ii} + \sum_{j \neq i} |a_{ij}| \leq -\varepsilon_i < 0, \quad i=1, 2, \dots, n \quad (9)$$

the velocity  $\dot{x}$  will be directed towards the interior of  $H$  at every vertex. Then, if the  $a_{ij}(t, x)$  satisfy (9) and (5) for all  $t \geq t_0$  and all  $x$ , the trivial solution  $x = 0$  is asymptotically and uniformly stable in the large. A Liapunov function is

$$V = \max_i |x_i| \quad (10)$$

For the third example consider a plant which can be described by the linear equation with constant coefficients

$$\dot{x} = Bx \quad (11)$$

For simplicity suppose that every root  $\lambda$  of  $|Bu + \lambda u| = 0$  is real, simple and positive. This is a common situation in some chemical engineering applications<sup>4</sup>. Let the plant be controlled in the manner suggested by the equation

$$\begin{pmatrix} \dot{\xi} \\ \dot{x} \end{pmatrix} = \begin{pmatrix} -r_0 & g'_0 \\ h_0 & B \end{pmatrix} \begin{pmatrix} \xi \\ x \end{pmatrix} \quad (12)$$

where  $r_0$ ,  $g_0$  and  $h_0$  are functions of  $t$ ,  $\xi$ ,  $x$ . Make a transformation of variables from  $\xi$ ,  $x$  to  $\zeta$ ,  $y$  in such a way that  $B$  is brought to diagonal form and

$$\begin{pmatrix} \dot{\zeta} \\ \dot{y}_1 \\ \dot{y}_2 \\ \vdots \\ \dot{y}_n \end{pmatrix} = \begin{pmatrix} -r & g_1 & g_2 & \dots & g_n \\ h_1 & -\lambda_1 & 0 & \dots & 0 \\ h_2 & 0 & -\lambda_2 & \dots & 0 \\ \dots & \dots & \dots & \dots & \dots \\ h_n & 0 & 0 & \dots & -\lambda_n \end{pmatrix} \begin{pmatrix} \zeta \\ y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix} \quad (13)$$

Here the  $r$ ,  $g_i$  and  $h_i$  are functions of  $t$ ,  $\zeta$ ,  $y$ .

Now take the vertices of  $H(1)$  at the  $4n$  points

$$\begin{aligned} &(\pm\alpha, \pm 1, 0, 0, \dots, 0) \\ &(\pm\alpha, 0, \pm 1, 0, \dots, 0) \\ &(\pm\alpha, 0, 0, \pm 1, \dots, 0) \\ &\dots\dots\dots \\ &(\pm\alpha, 0, 0, 0, \dots, \pm 1) \end{aligned} \quad (14)$$

At the  $2n$  vertices where  $\xi = \alpha > 0$ ,

$$\dot{\xi} = -r\alpha \pm g_i, \quad i=1, 2, \dots, n \quad (15)$$

which will be negative if

$$\left. \begin{aligned} \gamma - |g_i| &\geq \varepsilon_1 > 0, \quad i=1, 2, \dots, n \\ r &\geq \rho_1 > 0 \\ \rho_1 \alpha &\geq \gamma \end{aligned} \right\} \quad (16)$$

At the same points

$$\dot{y}_i = h_i \alpha - \lambda_i \operatorname{sgn} y_i \quad (17)$$

and  $(\dot{\xi}, \dot{y})'$  will be directed towards the interior of  $H$  at each of these points if in addition to (16)

$$\left. \begin{aligned} \lambda_i &\geq \mu_1 > 0, \quad i=1, 2, \dots, n \\ \eta - \sum_i |h_i| &\geq \varepsilon_2 > 0 \\ \mu_1 &\geq \alpha \eta \end{aligned} \right\} \quad (18)$$

The remaining  $2n$  vertices give no further condition. Now choose  $\alpha = \gamma/\rho_1$ , which satisfies the last of inequalities (16). The conditions

$$\left. \begin{aligned} |g_i| + \varepsilon_1 &\leq \gamma; \quad i=1, 2, \dots, n; \quad \varepsilon_1 > 0 \\ \sum_i |h_i| + \varepsilon_2 &\leq \eta; \quad \varepsilon_2 > 0 \\ 0 &< \rho_1 \leq r \leq \rho_2 \\ 0 &< \mu_1 \leq \lambda_i \leq \mu_2; \quad i=1, 2, \dots, n \\ \gamma \eta &\leq \mu_1 \rho_1 \end{aligned} \right\} \quad (19)$$

restrict the matrix in eqn (13) to a closed region  $G$  in  $E_{(n+1)^2}$ . If they are satisfied for all  $t \geq t_0$ ,  $\xi$ ,  $y$ , the solution  $\xi = y = 0$  is asymptotically and uniformly stable in the large.

The system corresponding to eqn (12) is shown in Figure 1. It is a generalization of one considered by Lurie<sup>5</sup> and by Letov<sup>6</sup>, their corresponding system being obtained by putting

$$g'_0(t, \xi, x) = r_0(t, \xi, x) \xi = f(c'x - k\xi) \quad (20)$$

$$h_0(t, \xi, x) = d \quad (21)$$

for some constant  $k$ ,  $c$ ,  $d$ . The important condition among inequalities (19) is the last, which shows that the system is always stable provided that  $r$  is sufficiently large for all  $t$ ,  $\xi$  and  $x$ . This agrees with the results of Lurie and Letov, but (contrary to a statement of LaSalle and Lefschetz<sup>7</sup>) larger values of  $r$  imply looser control.

Conditions (19) are in general more restrictive than those obtained by the methods of Lurie or Letov (where those apply).

For example, because  $H$  was chosen with a particular symmetry, conditions (19) make no distinction between positive and negative values of  $g_i$  or  $h_i$ . Systems stable according to (19) will therefore remain stable when the sign of the feedback around the plant is reversed.

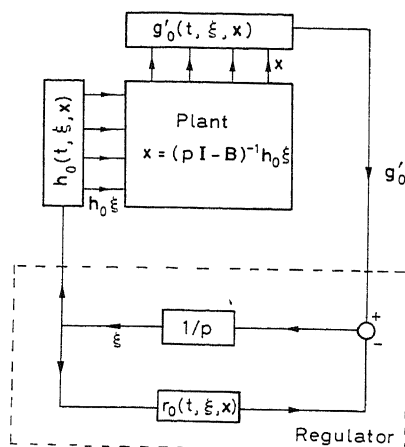


Figure 1

A more judicious choice of  $H$  would no doubt give better results at the cost of some increased complication. Such improved results would still be expected to give more stringent conditions than are obtained by Lurie and Letov from an initially more restricted class of systems.

### Comments

Stability criteria of the type considered restrict the elements of  $A$  by the condition  $A \in G$ , but leave the behaviour of the elements otherwise unrestricted. It is then possible for  $A$  to assume a constant value  $A_0 \in G$ . The necessary and sufficient condition for stability in this case is that every  $\text{Re } \lambda(A_0) \geq \varepsilon > 0$ , where  $\lambda$  is defined by

$$|A_0 u + \lambda u| = 0 \quad (22)$$

Thus it follows that the condition  $A \in G$  must ensure that every  $\text{Re } \lambda(A) \geq \varepsilon > 0$ . Inequality (4) (for all  $j$ ) and inequality (9)

are two known conditions of this type. Inequalities (19) represent a new result of the same kind which may be written, with reference to eqn (13),

$$r > \frac{(\max_i |g_{il}|) \sum_i |h_i|}{\min_i \lambda_i} > 0 \quad (23)$$

The theorem which has been proved may therefore be regarded, from one point of view, as a device for generating such conditions.

In many practical situations the equations of a system can be replaced in any one working condition by a linear autonomous approximation, but the appropriate linearization changes as the working conditions change. If the system is to be stable in each fixed working condition the  $\lambda$  must satisfy the criterion  $\text{Re } \lambda \geq \varepsilon > 0$ . It would be desirable to have additional criteria for the  $\lambda$  which ensured that the system was stable when the working conditions were changing. Such results can be obtained by the present method<sup>1</sup>, but only when  $A$  has a special form. Since  $A$  has  $n^2$  elements, conditions on the  $\lambda$  cannot generally restrict  $A$  to a closed region  $G$ : for this to be possible  $A$  must have no more than  $n$  independent variable elements.

### References

- ROSENBRICK, H. H. On the stability of a second-order differential equation. *J. Lond. math. Soc.* (to be published)
- ROSENBRICK, H. H. A Liapunov function with applications to some nonlinear physical systems. *Automatica* 1 (1963) 31
- ROSENBRICK, H. H. A Liapunov function for some naturally-occurring linear homogeneous time-dependent equations. *Automatica* 1 (1963)
- ROSENBRICK, H. H. The transient behaviour of distillation columns and heat exchangers. A historical and critical review. *Trans. Instn. chem. Engrs, Lond.* 40 (1962) 376
- LURIE, A. I. *Some Nonlinear Problems in the Theory of Automatic Control*. 1957. London; H. M. S. O.
- LETOV, A. M. *Stability in Nonlinear Control Systems*. 1961. Princeton; Princeton Univ. Press
- LASALLE, J. P. and LEFSCHETZ, S. *Stability by Liapunov's Direct Method*. p. 86. 1961. New York; Academic Press

### DISCUSSION

M. L. DEUTSCH, *Research Department, Socony Mobil Oil Company, Inc., Paulsboro, New Jersey, U.S.A.*

For this discussion, regions such as Rosenbrock's  $H$  are called 'starlike'. It appears that the crux of the matter is convexity rather than starlike character. For if we properly connect the vertices of  $H$  by hyperplanes, we produce a convex polyhedron,  $P$ , such that  $x \in H = x \in P$ , and to which the results of the paper apply. The application is then in some sense even more general, because the requirement of an inward-pointing  $\dot{x}$  at each of the vertices of  $P$  is less restrictive than the requirement of such a set of  $x$ 's at the vertices of  $H$  (Figure A).

At  $V$ , for example, if  $x$  lies in  $H$  it certainly lies in  $P$ , but not conversely. The region  $P$  thus defines a good Liapunov function.

Further, Rosenbrock's conditions do not guarantee stability in all starlike regions,  $H$ . Consider Figure B.

For this starlike region,  $H_1$ , the condition that  $\dot{x}$  points inward at  $V_1$  and  $V_2$  does not guarantee that  $\dot{x}$  points inward along the line  $V_1 V_2$ . But the line  $V_1 V_3$  converts the starlike region,  $H_1$ , into a convex polyhedron, to which Rosenbrock's conditions and results apply.

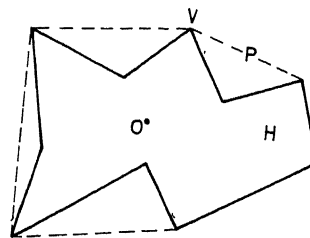


Figure A

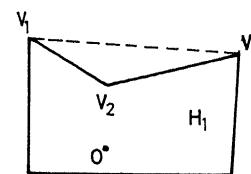


Figure B



H. H. ROSENBROCK, *in reply*

I am indebted to Dr. Deutsch for pointing out an error in the paper, which has since been corrected. The important condition in the proof is that  $n'\dot{x}$  should be positive on every face at each vertex of  $H$ . This condition was equated in the paper with the condition that  $n$  should be directed towards the interior of  $H$  at each corner. At a re-entrant corner this is not sufficient: the velocity must be directed out of the region exterior to  $H$ .

With the amendment the result in the paper is correct for star-shaped regions. This will give no further information about stability than the corresponding convex regions, and the latter will be simpler to handle and therefore generally preferable. There may, however, sometimes be an advantage in using a star-shaped region, as this will give more information about the trajectories than can be obtained from the convex region.

P. C. PARKS, *Department of Aeronautics, The University, Southampton, England*

There is a connection between the author's treatment of the system  $\dot{x} = Ax$  with his Liapunov function and the so-called Gershgorin theorem about the location of the latent roots of  $A$ . This theorem states that the latent roots are situated within the union of circles with centres  $a_{ii}$  and radii

$$\sum_{j=1}^n |a_{ij}| \quad i = 1, 2, \dots, n$$

(or alternatively, of radii

$$\sum_{j=1}^n |a_{ji}|).$$

The author's criterion, obtained by Liapunov's second method, specifies that one of these two unions of circles lies within the left-hand half of the complex plane.

H. H. ROSENBROCK, *in reply*

The theorem quoted by Mr. Parks does indeed give the 'known conditions' mentioned after eqn (22) of the paper.

M. MANSOUR, *Institut für Automatik und Industrielle Elektronik, Altbachstrasse 531, Dietlikon ZH, Switzerland*

My first remark deals with the statement in the second paragraph 'every ray from the origin intersects the boundary of  $H$  once and only once'. My idea is that it is not sufficient, and in order to prove the theorem this may be replaced by the statement 'any ray must not intersect the boundary of  $H$  more than twice' to avoid the possibility of  $n'\dot{x}$  to be negative at some points on the hyperplane. This possibility was also pointed out by Dr. Deutsch in his discussion.

My second remark is that the hypersphere

$$x_1^2 + x_2^2 + \dots + x_n^2 = r^2$$

can be considered as a limiting case of some closed region bounded by hyperplanes. The sufficient conditions for asymptotic stability of every solution of  $\dot{x} = A(t, x)x$  starting at time  $t_0$  from some  $x_0 \in H(r)$  is that at each point of the boundary

$$x_1^2 + x_2^2 + \dots + x_n^2 = r^2,$$

$\dot{x}$  is not zero and directed into the interior of  $H(r)$  for all  $A \in G$ . For a hypersphere this condition is equivalent to  $x'\dot{x}$  negative. i.e.

$$x_1\dot{x}_1 + x_2\dot{x}_2 + \dots + x_n\dot{x}_n < 0$$

Let  $A$  be the matrix

$$\begin{pmatrix} a_{11} & a_{12} & \dots & a_{1n} \\ \vdots & \vdots & & \vdots \\ a_{n1} & a_{n2} & & a_{nn} \end{pmatrix}$$

where  $a_{ij}$  is a function of  $t$  and  $x$   $i, j = 1 \dots n$

$$x^1 \dot{x} = x^1 \begin{pmatrix} a_{11} & \frac{a_{12} + a_{21}}{2} & \dots & \frac{a_{1n} + a_{n1}}{2} \\ \frac{a_{12} + a_{21}}{2} & a_{22} & & \vdots \\ \vdots & & & \vdots \\ \frac{a_{1n} + a_{n1}}{2} & & & a_{nn} \end{pmatrix} x = -x^1 C x$$

This is a quadratic form whose matrix is  $C$ . For  $x^1 x^1$  to be negative definite the matrix  $C$  must satisfy the Sylvester inequalities

$$-a_{11} > 0 \quad \begin{vmatrix} -a_{11} & -\frac{a_{12} + a_{21}}{2} \\ -\frac{a_{12} + a_{21}}{2} & -a_{22} \end{vmatrix} > 0$$

$$\begin{vmatrix} -a_{11} & -\left(\frac{a_{12} + a_{21}}{2}\right) & -\left(\frac{a_{13} + a_{31}}{2}\right) \\ -\left(\frac{a_{12} + a_{21}}{2}\right) & -a_{22} & -\left(\frac{a_{23} + a_{32}}{2}\right) \\ -\left(\frac{a_{13} + a_{31}}{2}\right) & -\left(\frac{a_{23} + a_{32}}{2}\right) & -a_{33} \end{vmatrix} > 0 \dots \text{etc.}$$

These are sufficient conditions for the stability of the considered system. The above is equivalent to taking a Liapunov function

$$V = x^1 E x \quad \text{where } E \text{ is the unit matrix} \\ \dot{V} = -2 x^1 C x \quad \text{where } -2 C = A + A^1$$

Applying this to the third example in the paper gives

$$r > \frac{1}{4} \sum_i \frac{(g_i + h_i)^2}{\lambda_i} > 0$$

This condition seems to be more flexible and gives better results, especially when  $g_i + h_i$  cancel to some extent. However, this is not always the case.

H. H. ROSENBROCK, *in reply*

The condition suggested by Dr. Mansour is that any straight line not in the boundary of  $H$  should, at most, intersect it twice. This is equivalent to convexity. It would also be necessary to ensure that  $H$  enclosed the origin. With this change in the conditions of  $H$  the original wording of the paper, mentioned in the reply to Dr. Deutsch, would be correct.

The relative efficiency of different types of Liapunov functions depends on the problem considered. The result (19) of the paper, as stated therein, is not a good one. It could be improved by a better choice of  $H$ , but might still be worse than the result obtained by Dr. Mansour, using Krasovskii's method. The result in (4), on the other hand, is better for many physical systems than results obtained by Krasovskii's method.

## GENERAL DISCUSSION OF STABILITY PROBLEMS

J. P. LASALLE

We all agree that the only method available for studying the stability of non-linear systems, which effectively takes into account the non-linearities of the system, is Liapunov's second method. We have, however, a great deal of theory, and many special problems have been solved. I think that the most important problem for today is to find a means of making use of the large-scale computers which are available for applying Liapunov's second method to control problems in a more effective way.

I personally do not see any sign that we are close to a solution of this problem.

G. P. SZEGÖ

I certainly agree with Dr. LaSalle that the solution of the stability problem of control systems lies in the use of digital computers for constructing Liapunov functions which describe the behaviour of the system in the whole space.

I would like, first of all, to justify the use of a computer for this particular task. It has been said that with a computer one can, after all, integrate the equations. Unfortunately one would have to integrate the differential equation for all initial conditions. On the other hand, for a Liapunov function which defines the behaviour of the system in the whole space the computation will be performed once and only once.

There are mainly two different ways of approaching this problem. One is to solve a partial differential equation of the Zubov type or one of the partial differential equations I have presented in this section which will cover a wider range of problems. Of course, in this approach

one has to make the assumption that there exists a polynomial Liapunov function which satisfies the partial differential equation.

Assume now that we are able to do this, or better, that the computer is able to do this: the stability problem is not yet solved. The computer will give a polynomial form in  $n$  variables of order  $m$ . The problem is now reduced to the analysis of this form in order to decide what are the stability properties of the system. This task is not a trivial one.

Another way of attacking this problem is to follow the method presented by Dr. Nesbit in a previous session. This method, however, also presents some unanswered questions in the case in which the non-linearities depend upon many components of the state vector. Both approaches seem to be feasible; the choice of one or the other strictly depends upon the particular problem in question. If one has to investigate a first system, or a well-defined equation, one has to use the first approach. If one is interested in the relative stability properties of one system with respect to the neighbouring ones, very likely the second approach will be more convenient.

H. H. ROSENBROCK

I agree with Dr. LaSalle that it would be most attractive to use a digital computer to generate Liapunov functions. The method which I thought of trying was to specify the corners of  $H$  and allow the computer to examine each face at each corner. It would have to do this for all matrices in  $G$ , but this would be feasible if  $G$  were, say, a line. The computer would probably find that the velocity was directed out of  $H$  at one or more corners. Then it would be necessary to have some rules by which the computer could move the corners in order to improve the position. One might also perhaps allow the computer to put in fresh corners.

# SYSTEM DYNAMICS AND OTHER PROBLEMS

---

## Process Dynamics and its Applications to Industrial Process Design and Process Control

A Survey by T. J. WILLIAMS

### Introduction

In introducing the subject of process dynamics, one should perhaps first define the area to be covered. While process dynamics has been used by others to cover a wider area, I would like to confine my remarks to topics related to the so-called process industries. That is, such industries as chemicals, petroleum, glass, cement, metals, food, and related industries where chemical and mechanical changes are brought about in a continuous or semicontinuous stream of raw materials in order to produce a product or products whose bulk chemical or physical properties are the important source of their economic value. These remarks will also apply generally to much of the energy conversion industries who transform chemical or potential energy into the more generally usable electrical and thermal forms. These remarks will not apply generally to those manufacturing industries which produce a class of discrete objects as products, the individual properties of which objects are the source of their economic value.

The study of process dynamics can serve as a source of intellectual satisfaction for the new knowledge obtained. It can also provide the possible remedy of an existing plant control problem. However, the major portion of all of our studies of process dynamics, or of the time dependent responses of our processes, should have as its ultimate purpose that of aiding the future design of new and/or improved control systems for our processes or of aiding the future design or operation of these processes themselves.

The basic need in plant engineering is for sufficient knowledge of industrial processing plants, their manufacturing tools and machinery, and their related distribution and economic systems, so that one can theoretically characterize and compute a satisfactory representation of their dynamic behaviour in the face of any postulated external or internal influence. In addition, one must be able to do this in the design stage of a new plant or while a new component design is still on the drawing board.

Now and in the interim before this ultimate is attained there is, of course, much need for retrofitting of the present plants so that they can produce to their optimum capacity and at the highest possible profits. However, such work is necessitated by

deficiencies in our past knowledge and abilities in design. Such deficiencies must and should decrease as the overall design and prediction ability increases.

The ultimate future success, therefore, depends upon our ability to produce mathematical models of the plants and of their related devices and processes. The ability to do this for time-varying requirements and for time-dependent upsets depends upon the knowledge of process dynamics. While traditionally the engineering of production plants has been on a 'steady-state' or 'smooth-operation' basis, allowing the use of algebraic equation models, there is now no serious question but that future design must be on a dynamic or unsteady-state basis and must therefore involve differential equations or related models. Again process dynamics knowledge makes this possible.

It is therefore important to use this opportunity in order to review the recent progress in the field of process dynamics and its applications to industrial process control in the above light. This is an excellent occasion to establish the present status of the field, at least as viewed by this author, and to make some recommendations for future work which may serve to further the field and remedy some of the omissions noted at this time.

It will be seen that future progress in process dynamics studies and the application of this knowledge to process control and process design problems depends on developments in three areas. These are: better methods of approximating complex responses by simple models; more complete and more exact characterizations of the non-linear parameters of process systems; and the development and use of faster and more capable computer systems for process simulation use.

The reader's attention is also invited to the several review articles and books covering the field of process dynamics<sup>3, 31, 38-41</sup>, particularly the subject surveys of Archer and Rothfus<sup>2</sup> on distillation dynamics and control, Lapidus<sup>20</sup> on chemical reactor dynamics; and Williams and Morris<sup>42</sup> on heat exchanger responses. The report of the American Automatic Control Council edited by Berger<sup>8</sup> also contains much of value to the subject covered here.

## Facts about Process Industry Systems

In order to set the stage properly for a discussion of the mathematical modelling of process industry manufacturing systems, one should call attention to the important characteristics of such systems which distinguish them from those of other areas. *Table 1* lists several such factors. They serve to emphasize a statement which is especially true for these industries: 'The real world is non-linear, multivariable, and non-stochastic—all the things which make a theoretical analysis difficult'.

## The Uses of Process Dynamics in Process Plant Studies

### *A Division of Objectives and of Approach*

At this time, process dynamics studies are being used for several purposes in process plant applications. However, these are such that they require only two main outlooks in developing the process dynamics concerned. These are:

(1) Development of overall approximate models of process dynamics either from experimental work or from drastically lumped and/or linearized forms of theoretical plant mathematical models. The object here is to produce the simplest possible model which still adequately represents the phenomena studied or which provides the response required for a specific purpose. Some examples are:

(a) Feedforward and feedback compensation in plant control systems, particularly analogue computer control systems<sup>10, 21, 29</sup>.

(b) Linearized simulations for classroom demonstrations or control systems development work<sup>29</sup>.

(2) Studies of plant design problems and simulations of plant start-up and major upset conditions. Here, in contrast to the above, the intention is to obtain as complete and as exact a duplication of plant behaviour as is possible. In fact, the limit of the study is usually the size of the analogue computing equipment available or the amount of digital computer calculation time which can be used<sup>40</sup>.

These two outlooks have, of course, entirely different requirements for simulation accuracy, variable excursion range, treatment of linearizations, etc. In addition as computers become ever more available and more capable, the disparity in approach used and results obtained from these two types of studies will necessarily increase. Both have a definite place. They are mentioned here only to call attention to the fact, mentioned later, that care must be taken in comparing the results of various studies to note the purpose for which they were developed.

An excellent comparison of the amount of model complexity necessary for various applications of process dynamics is provided by the generalized chemical reactor model of Williams and Otto<sup>48</sup> and the several solutions proposed for it<sup>7, 10, 12, 13, 17</sup>.

Eventually one must be able to develop and use dynamic models for complete plants of all types. As an example of such work, Borel<sup>11</sup> has developed for this Congress his paper on the regulation of a hydro-electric plant.

### *Future Gains from Process Dynamics Knowledge in Process Design and Control Work*

The gains which process dynamics knowledge can make to process control work are already well known from many sources. In addition, a knowledge of process dynamics offers the possibility of making the following gains in plant design and opera-

tion available to justify the time and expense necessary for its development:

(1) In many plant situations the use of a continuously dynamic situation rather than a steady-state method of operation will result in increased gains from a process such as in some batch reactions compared to continuous ones and pulsed extraction columns *versus* steady operation of the same columns. Process dynamics analyses can evaluate such gains and determine their economic worth.

(2) Process dynamics knowledge allows the determination of the true nature of the operating extremes to be expected from a process under various conditions of upset. It can thus help to avoid excess costs arising from overdesign of the process and/or its control system. Likewise, it allows the determination of the amount of time off-specification following an upset and the equivalent amount of off-specification product formed.

(3) Process dynamics information can supplement theoretical reasoning and steady-state investigations in elucidating the nature of some processes, thus helping to establish models useful for design as well as for control.

## The Formulation and Use of Theoretical Models

It is well recognized that one can at present write theoretical mathematical models for practically all production processes to whatever degree of complexity one is capable of solving on the available computing equipment. That is, one already has a basic knowledge of the probable behaviour of processing systems and can list the parameters that will be important. However, this is only a minor part of our difficulties and there are still many problems requiring investigation. Some of these are:

(1) One has at present only a rudimentary knowledge of the amount of detail in a mathematical model which is actually necessary to simulate the dynamic behaviour of a particular process to achieve a desired objective. This is particularly true in terms of approximate models and is due to several factors:

(a) The tremendous range of complexity of the processes themselves as mentioned in *Table 1*.

(b) The fact that there is no generally accepted definition of the goodness of fit required between experimental and theoretical or simulation data.

(c) Since simulations can be used for many purposes, this in itself allows a tremendous breadth in the requirements for modelling detail and the resulting required closeness of fit between theoretical and experimental data.

(d) Deficiencies in literature examples of the use of theoretical process models. In many cases severe limitations in computing capacity available to investigators have put too restrictive limits on the modelling detail which was actually used.

(e) Process dynamics studies to date have tended to be generally either all experimental or all theoretical with little comparison of one type of data with the other for a particular piece of apparatus.

(2) A great deal of work has been done recently on the linearization of the mathematical models of processes through the use of the technique of small perturbations. This has resulted in an ability to simulate much larger and more complex pieces of equipment and systems. However, this has itself raised several problems related to those just above.

Table 1. Vital Factors Concerning the Process Industries Important to a Discussion of Mathematical Modelling

- I These industries processes or separate operations are, as a rule, multistage, multiphase, and multicomponent and operate over wide ranges of temperature, pressure, and flow. In addition, they often must be treated as distributed parameter systems and usually as non-linear ones.
- II A typical production plant will usually consist of from several to many such processes connected together to form a complete unit.
- III The so-called 'time constants' of the several important variables in a process may typically vary by one to several orders of magnitude. This is particularly true when one is comparing temperature, composition and flow transients for the same process.
- IV As a general rule, these processes will be operated below capacity because of competition with other processes or even other completely different products. Thus gains from optimizations must often be limited only to increased yields or lowered costs rather than being based on higher production levels from the plant.
- V Because of the rapidly moving pace of industrial research and the vagaries of the economy, management is usually anxious to rush a new plant into production as soon as possible after approval of its construction. There is thus a supreme premium on time during the design and construction phases.
- VI The design and operating details between competing processes may vary slightly and still result in appreciable competitive advantages to one process or the other. For this reason there is a reluctance to discuss actual operating process details or real research data in the literature. At the same time discovery of new but comparably small improvements in such operating methods or designs can be equally important.

(a) What are the actual limitations on the linearization techniques used, i.e. through what ranges of the variables involved can the linearized simulation be perturbed and still duplicate the actual processes to some specified closeness of fit?

(b) Perhaps some process variables are more susceptible to linearization than others and a mixed linear—non-linear system is the proper answer.

The above statements are particularly important when one is considering process control studies similar to Item (1) under the heading 'Uses of Process Dynamics' given above. These difficulties are also of extreme importance as long as computing capacity limitations restrict the detail possible in a plant design or operation study.

### Experimental Methods for Determining Process Dynamics

Activity in the experimental testing of process plants and equipment to determine the overall dynamics has been very active during the past few years. Chief interest has been concentrated in the use of pulse testing methods<sup>15, 18</sup> with subsequent computer data reduction to frequency response or in the use of correlation methods. These latter have often used special computers<sup>9</sup> built specifically for this purpose. These automatic data-reduction techniques with computers have made it relatively easy to develop overall transfer function or performance function models relating a particular process output variable to a particular input variable change. This is especially true for pulse type inputs. Because of the special data reduction problems involved, however, higher-order effects may be lost or are at best very difficult to obtain. The following questions regarding these studies remain relatively unanswered as of this time:

(1) What are the indispensable dependent variables of any particular process? That is, just how many separate transfer functions are necessary to describe the system adequately? What are the best data-presentation methods for expressing this data?

(2) Just as in the case of linearized small-perturbation type simulations what are the limits of input variable manipulation for which the linear transfer function representation is adequate?

(3) What are the criteria to be followed in deciding whether higher-order effects may or may not be important in any specific case?

### Status of Process Dynamics Studies for Various Process Operations

Studies of process dynamics work to date have been concentrated mainly in three areas: heat exchangers, distillation columns, and chemical reactors. Heat exchangers have been very important because of their ready availability in nearly all universities and the fact that they can be operated over wide ranges of operation with only steam and water, both of which are readily available in all laboratories. In addition, they provide intriguing experimental models for investigating the all-important problem of lumped parameter *versus* distributed parameter representation<sup>6, 14, 16, 25, 35, 42</sup>. Distillation columns are excellent examples of multivariable systems and lend themselves to many types of representation<sup>1, 2, 4, 5, 19, 21-24, 29, 32, 33, 44</sup>. Also, they are extremely important commercially so that solutions to their responses and control problems are very valuable economically. Likewise, chemical reactors are very important commercially being the heart of any chemical process. They also are intriguing problems in process optimizations because of the complexities and the interactions taking place in chemical reactions and the presence of recycle streams in most cases<sup>3, 7, 10, 12, 17, 20, 28, 36, 37, 39</sup>.

Based on this one can make several statements on further work which should be carried out in these subjects. These are given below:

(1) Heat exchangers—A very large amount of work has been done in this field, so much so that additional work would have marginal value unless devoted to one of the following topics or others of related complexity:

(a) Dynamics of fluid behaviour in the shell section of shell and tube exchangers and the effect of these flow changes upon tube side temperatures.

(b) Dynamics of forced circulation and thermosiphon reboilers, condensers, and other such heat transfer devices where a phase change in a flowing stream is of major importance. The paper of M'Pherson and Muscettola<sup>25</sup> treats this specific problem in the context of a nuclear reactor. The results will be valuable for ordinary boilers as well.

(2) Distillation—While distillation has been a popular subject for study in the past, major problems exist in the following areas:

(a) What are the valid approximate models which may be devised to represent the distillation system since the complete mathematical model of a 'real' column is completely unmanageable by present methods? What are the limitations of each type of approximation that may be devised? In this regard, it should be noted that several of the papers of this Congress have concern-

ed themselves with the development of approximate models for distillation columns. I refer specifically to those of Izawa and Morinaga<sup>19</sup>, Moczek, Otto and Williams<sup>24</sup>, and Zavorka<sup>44</sup>.

(b) Characterization of plate mixing and mass transfer effects necessary in the more exact simulations. In this Congress the vital nature of this subject is recognized by the reports of Anisimov<sup>1</sup> who discusses plate mass transfer, and by Takamutsu and Nakanishi<sup>34</sup> who discuss both topics.

(3) Reactors—The simpler reactor systems have been very extensively studied for the cases of relatively simple kinetic situations. Much work needs to be done yet on:

(a) Complex reactor designs. (The paper of van der Heyden and van Nes<sup>36</sup> of this Congress discusses a very real problem in reactor design and control—the heat exchange system of a catalytic cracker<sup>36</sup>.)

(b) Higher order and multiple reaction kinetic systems. (The paper of Volter<sup>37</sup> of this Congress treats this extremely complex and important subject.)

(c) Optimization (dynamic and steady state) and adaptive control of 'real' reactor situations. Reactor design and operation is the area where optimization and adaptation appears to hold the greatest promise.

(4) Others—Relatively little process dynamics and control work has been done for such chemical process operations as multiple effect evaporators, absorption columns, extraction vessels and columns, continuous filtration, drying and humidification, etc. Work is under way at several places on extraction which may well place this operation on a par with distillation in a year or so. While work has mainly been concentrated in the more common chemical and petroleum industry examples, it should be remembered that cement kilns, glass furnaces, blast furnaces, and smelters are all chemical reactors and the findings on the study of chemical reactors will eventually be translatable to these more complex cases.

One place where much more work is necessary and where additional effort should be very rewarding is in the study and characterization of plastic flow, particularly in the extrusion of metals and plastics and the rolling and cold working of metals.

#### General Notes Regarding Process Dynamics

Some general observations regarding the problem areas in process dynamics can also be made at this time. These are:

(1) Most difficulties in mathematical model-making resolve themselves directly to the characterization of fluid flow phenomena occurring in the process such as turbulence, liquid mixing, etc. In this respect the important work of Nichols<sup>26</sup>, and Oldenburger and Goodson<sup>27</sup> on the dynamics of hydraulic and pneumatic control lines is extremely valuable. In addition, the work of Delvaux<sup>14</sup> and Takamutsu and Nakanishi<sup>34</sup> on fluid mixing in heat exchangers and on distillation column trays gives important insights into the solutions to this problem.

(2) Control in chemical processing almost always involves the manipulation of the flow of a fluid stream (here fluid may mean gas, liquid, or fluidized solid).

Hence fluid flow problems are at the heart of all control investigations, while at the same time being the process dynamics phenomena which one least understands.

#### The Place of the Computer in Process Dynamics Studies

As mentioned earlier, the availability of computer capacity and speed of operation is one of the main deterrents to further progress in the application of process dynamics studies to plant operation and design problems. It will be helpful to establish the present status of such computer usage before proceeding.

#### Analogues Versus Digitals for Process Simulations

A point worth noting and commenting upon is the use of analogue versus digital computers for simulation problems, i. e. solution of ordinary and partial differential equations for the study of a transient process behaviour or of an automatic control response. The real criterion of applicability here is whether or not the computer can operate in 'real time'. That is, can it compute sufficiently fast so that its answers appear faster or at least at the same rate at which the physical system being simulated would have produced them. Therefore, both the complexity and speed of response of the system being studied is important. In Figure 1 this is expressed as the order of the system (approximately equal to the number of separate differential equations

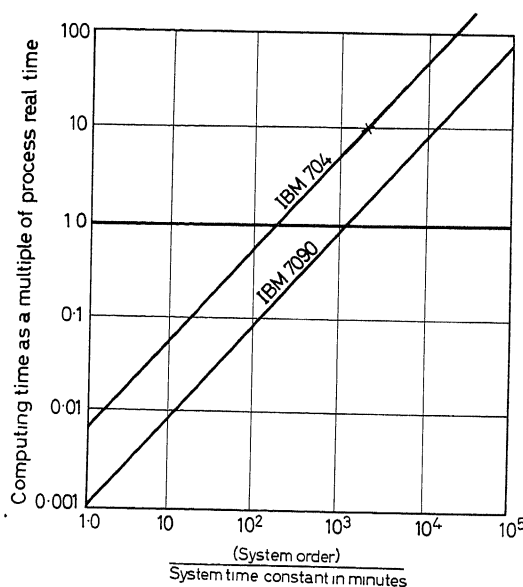


Figure 1. An evaluation of digital computers as process simulators as a function of systems order and response speed (data are based on computation of distillation column dynamics)

in the system) divided by the time constant of each cell (the time required for the cell output to respond 63.2 per cent to a step input change). Under such a criterion, Figure 1 shows the capability of the IBM 704 and IBM 7090 to compute the transient response of a relatively complex, high-order, system, here a multicomponent distillation column.

The indicated point is for a particular column studied<sup>24</sup>. In this case the IBM 704 was able to compute only  $1/10$  as fast as real time. In order to achieve column response speeds a reduction of system order, i. e. number of plates, to  $1/10$  of the original number or of the system complexity, i. e. detail of plate equilibrium and flow computations, by the same amount would be necessary.

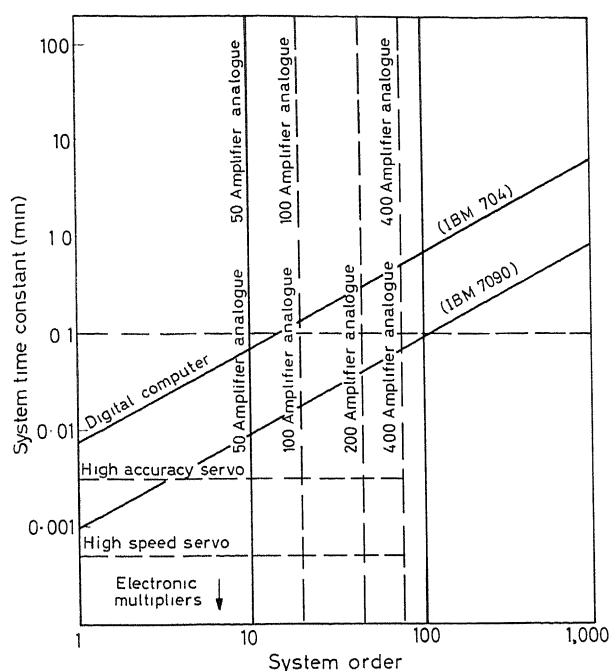


Figure 2(a). Diagram of relative applicability of analogue and digital computers to process simulation problems

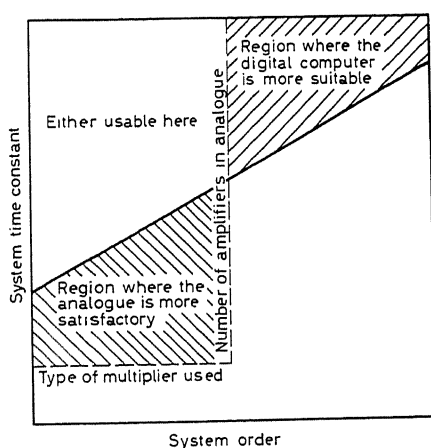


Figure 2(b). Explanation of data in Figure 2(a)

Figure 2 shows the relationship of various sizes and types of analogue computer installations to the 704 and 7090 for simulation purposes. Analogue computers are limited directly in the system order which they can duplicate by the number of computing elements they contain and indirectly by the accuracy of these elements. They are limited in speed of response by the type of multiplying gear used. Thus on Figure 2(a), a particular analogue computer would be applicable to all combinations of order and of time constant above a horizontal line representing the type of multiplier used and to the left of a vertical line representing the number of amplifiers it contains. The digital can adequately handle all combinations above the line representing it on the graph since it is limited theoretically only by its speed of computation or of information handling. Figure 2(b) extends

Figure 2(a) by pointing out the regions on the graph where each machine would be superior to the other for process simulations.

It is truly unfortunate to note the number of chemical process systems which fall into the lower right-hand corner of Figure 2(b). That is, they are too large for the analogue machine and too fast for the digital computers at present available.

#### Computer Size and Convenience

Despite the tremendous advances which have already been made in the use of computers in the process industries, it is still disappointing to the author to see the low order of complexity of the problems which can be handled with the computing facilities available to most of our process control and systems engineers. This is particularly so when one reflects upon the tremendously high order of the equations required to duplicate the action of even a relatively small and comparatively simple chemical plant. Thus there is no doubt that one will always be able to use still larger and faster computers as they become available up to the limit which particular companies can invest in such work.

Reflection would lead one to judge that this limit could easily be as high as 0.5 per cent of the total invested capital of a progressive and growing chemical or petroleum company not counting those computers used strictly for accounting purposes. It is hoped that equally large and capable computers can be provided for the universities.

Necessary as the above growth in size and speed is, another factor is of even more importance for present and future growth. This is the question of accessibility of the machines to the average engineer. Methods must be developed to make it easy for any engineer who so desires to present his problem to the machine and to obtain his answers quickly and in a readily usable form. Unfortunately all computer work still requires the attention at some stage or other of the 'expert'. The much touted 'assembler' and 'compiler' routines have indeed made the task of programming much easier. However, even they require the assistance of an expert for their 'debugging'. Also, only the expert computer man can get the most from the machine when memory space or computing time is in short supply even with the aid of the automatic programming aids available at present. Thus much work yet needs to be done on such computer 'software'.

Physical location is also a factor in computer use. In many cases a suitable computer problem is done by a less accurate method or not done at all because the computer was not conveniently near to the engineer to use.

#### The Simulation of Fast, High-order Chemical Systems

Thus, one of the very serious drawbacks in present day process control and process design practice is our inability to simulate large, fast chemical systems because present day analogue computers are too small and digital computers too slow. Unfortunately, also, available analogue computer accuracies prevent these machines becoming much larger than the largest of those in use today (about 400 amplifiers) when applied to chemical problems.

However, there are several recent or imminent developments in analogue computer equipment and in operating techniques which promise to greatly aid and possibly even revolutionize analogue computer operation. Most of these involve digital techniques of one type or another.



One of the major present-day problems with analogue computer operation in the process industries is the difficulty of adequately representing a true time delay. The development of a suitable and reasonably priced magnetic tape or drum mechanism for this purpose would be a major boon to the field. Such devices are now under development.

As the reader knows, analogue computers are especially designed for the solution of ordinary differential equations and do this extremely well while digital computers are very fast and accurate in the solution of algebraic equations but relatively slow with large systems of differential equations. In addition, analogues have only rudimentary logic capabilities while extremely complex logic operations are carried out directly on digital machines. There is thus a very fruitful potential in the joint use of these machines for the solution of large process problems such as being discussed here, which entail all three types of mathematical operation, solution of algebraic equations, and logical selection from several possible paths of action. Devices for connecting the two computers together during operation are in use by several of the defense system contractors, such as Space Technology Laboratories, Convair Astronautics, and International Business Machines Company in the United States. Models of these devices are also under development for use by the process industries.

Another area where aid is needed in analogue computation is in the simulation of distributed parameter systems and processes. Adequate mathematical representation of these systems requires the use of partial differential equations or extremely large sets of simultaneous ordinary differential equations. Thus their solution on analogue equipment requires a very large machine. However, the use of a modified form of the data storage drum mentioned above for true time delay simulation offers an aid in the solution of this type of problem. If preliminary results can be stored during the solution of a problem of this type on the analogue computer, it can be solved on a much smaller machine than required above by the use of a repetitive trial and error procedure which can be readily programmed. Again, these devices are under active development by the computer manufacturers. Thus, digital devices promise to have a major effect on analogue computer operation in the process industries.

Now consider a possible set of requirements for such a connecting device as described above for use in solving industrial process simulation problems.

(1) In order to handle reactor problems with their associated large number of composition values, the device must be able to store the instantaneous values of at least ten variables at each stage of the computations. These may be compositions, temperatures, flows, etc.

(2) For time delay representation purposes, it should be able to hold values of each variable for at least 2 min.

(3) It should be able to give good resolution for frequencies as high as 10 c/sec. Thus sampling at a rate of at least 40 samples/sec/variable would be necessary.

(4) Combined with Item (2) above, this means that up to 4,800 separate numerical values per variable must be stored for an adequate time delay representation.

(5) Combination of Items (4) and (1) requires that a storage capability of at least 50,000 words be available for use with the analogue computer.

(6) In order to allow adequate communication between the

two machines (analogue and digital) at least ten channels each way will be necessary. Multiplexed analogue-digital and digital-analogue converters may be used if their speed is adequate for the task outlined above.

With such a combined analogue-digital computer the analogue computer could perform all integrations and such operations which are slow on the digital machine while the latter performed the arithmetic operations and function generation duties which require large amounts of equipment on the analogue. In this way a large inroad into the lower right-hand area of Figure 2(b) could be possible.

## References

- ANISIMOV, E. V. The studies of dynamics and statics of rectification process. *Automatic and Remote Control* 1963. London; Butterworths: Munich; Oldenbourg
- ARCHER, D. H. and ROTHFUS, R. R. The dynamics and control of distillation units and other mass transfer equipment. *Chem. Engng Progr. Symp. Ser.* 57, No. 36 (1961) 2
- ARIS, R. and AMUNDSON, N. R. An analysis of chemical reactor stability and control. *Chem. Engng Sci.* 7, No. 3 (1958) 121
- ARMSTRONG, W. D. and WOOD, R. M. The dynamic response of a distillation column to changes in the reflux and vapour flow rates. *Trans. Instn chem. Engrs, Lond.* 39 (1961) 65
- ARMSTRONG, W. D. and WOOD, R. M. An introduction to the theoretical evaluation of the frequency response of a distillation column to a change in reflux flow rate. *Trans. Instn chem. Engrs, Lond.* 39 (1961) 80
- BALL, S. J. Approximate models for distributed-parameter heat-transfer systems. Paper presented at the Fourth Joint Automatic Control Conference, University of Minnesota, Minneapolis, Minnesota, June 19-21, 1963
- BEECHER, A. E. and GOULD, L. A. Considerations in the design of a dynamic control system for the generalized chemical processing model considered as a nonlinear system. *Amer. Inst. elect. Engrs Spec. Publ. S 132, Computers in Control*, 146-156 (Sept. 1961)
- BERGER, A. R. Ed. *State of Automatic Control Research*, Report of Workshop on Automatic Control Research, National Science Foundation, Washington, D.C., 1962
- BLANDHOL, E. On the use of adjustable models for determining system dynamics, *Tech. Rep. No. 62-5-D*, Institutt for Regulerings-teknikk, Norges Tekniske Høgskole, Trondheim, Norway, March 1962
- BOLLINGER, R. E. and LAMB, D. E. The design of a combined feedforward-feedback control system. Paper presented at the Fourth Joint Automatic Control Conference, University of Minnesota, Minneapolis, Minnesota, June 1963
- BOREL, L. Equations of regulation of a hydro-electric installation. *Automatic and Remote Control* 1963. London; Butterworths: Munich; Oldenbourg
- BOYDSTON, R. E. A dynamic solution to a generalized chemical processing model. *Amer. Inst. elect. Engrs Spec. Publ., S 132, Computers in Control* 16-30 (September 1961)
- CHAPIN, D. W. A pneumatic computer for process control. *J. Instrum. Soc. Amer.* 8, No. 9 (1961) 38; No. 10 (1961) 53
- DELVAUX, L. Étude expérimentale du comportement dynamique d'un échangeur de chaleur et d'un processus de mélange. *Automatic and Remote Control* 1963. London; Butterworths: Munich; Oldenbourg
- DREIFKE, G. E. Effects of Input Pulse Shape and Width on Accuracy of Dynamic Systems Analysis With Experimental Pulse Data, *Sc.D. Dissertation*, Washington Univ. St. Louis, Missouri, June 1961
- ENNS, M. Comparison of dynamic models of a superheater, *Trans. Amer. Soc. mech. Engrs, Ser. C., J. Heat Transfer*, 84, No. 4 (1962) 375



- <sup>17</sup> GOULD, L. A. and KIPINIAK, W. Dynamic Optimization and Control of a Stirred Tank Chemical Reactor, *Amer. Inst. elect. Engrs, Spec. Publ., S 132, Computers in Control*, 229-241 (September 1961)
- <sup>18</sup> HOUGEN, J. O. and WALSH, R. A. Pulse Testing Method. *Chem. Engng Progr.* 57, No. 3 (1961) 69
- <sup>19</sup> IZAWA, K. and MORINAGA, T. Dynamic characteristics of binary distillation column. *Automatic and Remote Control* 1963. London; Butterworths: Munich; Oldenbourg
- <sup>20</sup> LAPIDUS, L. On the dynamics of chemical reactors. *Chem. Engng Progr. Symp. Ser.* 57, No. 36 (1961) 34
- <sup>21</sup> LUPFER, D. E. and JOHNSON, M. L. Automatic control of distillation columns to achieve optimum operation, Paper presented at the Fourth Joint Automatic Control Conference, University of Minnesota, Minneapolis, Minnesota, June 19-21, 1963
- <sup>22</sup> LUPFER, D. E. and OGLESBY, M. W. Automatic control of distillation columns. *Industr. Engng Chem. (Industr.)* 53 (1961) 963
- <sup>23</sup> LUPFER, D. E. and PARSONS, J. R. A predictive control system for distillation columns. *Chem. Engng Progr.* 58, No. 9 (1962) 37
- <sup>24</sup> MOCZEK, J. S., OTTO, R. E. and WILLIAMS, T. J. Approximation models for the dynamic response of large distillation columns. *Automatic and Remote Control* 1963. London; Butterworths: Munich; Oldenbourg
- <sup>25</sup> M'PHERSON, P. K. and MUSCETTOLA, M. A study of the dynamics of steam voids in boiling water nuclear reactors. *Automatic and Remote Control* 1963. London; Butterworths: Munich; Oldenbourg
- <sup>26</sup> NICHOLS, N. B. The linear properties of pneumatic transmission lines. *Trans. Instrum. Soc. Amer.* 1, No. 1 (1962) 5
- <sup>27</sup> OLDENBURGER, R. and GOODSON, R. E. Hydraulic line dynamics. *Automatic and Remote Control* 1963. London; Butterworths: Munich; Oldenbourg
- <sup>28</sup> PETERSEN, T. I. Reaction kinetics optimization using nonlinear estimation. *Chem. Engng Progr. Symp. Ser.* 56, No. 31 (1960) 111
- <sup>29</sup> RIPPIN, D. W. T. and LAMB, D. E. A theoretical study of the dynamics and control of binary distillation. Paper presented at the 53rd Annual Meeting, Amer. Inst. Chem. Engrs, Washington, D.C., December 1960
- <sup>30</sup> ROSE, A., JOHNSON, R. C., HEINY, R. L. and WILLIAMS, T. J. Computers, mathematics, statistics and automation, *Industr. Engng Chem. (Industr.)* 48 (1956) 622
- <sup>31</sup> ROSE, A., JOHNSON, R. C., HEINY, R. L., SCHILK, J. A. and WILLIAMS, T. J. Computers, mathematics, statistics, and automation. *Industr. Engng Chem. (Industr.)* 49, No. 3, Pt. II (1957) 554
- <sup>32</sup> ROSENBRICK, H. H. A theorem of 'Dynamic Conservation' for distillation. *Trans. Instn Chem. Engrs* 38 (1960) 279
- <sup>33</sup> ROSENBRICK, H. H. The control of distillation columns. *Trans. Instn Chem. Engrs* 40 (1962) 35
- <sup>34</sup> TAKAMUTSU, T. and NAKANISHI, E. Effects of fluid mixing and its expressions on dynamics of mass transfer process. *Automatic and Remote Control* 1963. London; Butterworths: Munich; Oldenbourg
- <sup>35</sup> THAL-LARSEN, H. Dynamics of heat exchangers and their models. *Trans. Amer. Soc. mech. Engrs, Ser. D, J. Basic Engng* 82 (1960) 489
- <sup>36</sup> VAN DER HEYDEN, C. A. J. M. and VAN NES, A. G. A study of dynamic behaviour of a catalytic cracker power-recovery system by means of an analogue computer. *Automatic and Remote Control* 1963. London; Butterworths: Munich; Oldenbourg
- <sup>37</sup> VOLTER, B. V. Automation of polyethylene production under high pressure. *Automatic and Remote Control* 1963. London; Butterworths: Munich; Oldenbourg
- <sup>38</sup> WILLIAMS, T. J. Process Control and Automation. *Industr. Engng Chem. (Industr.)* 50, No. 3, Pt. II (1958) 520; 51, No. 3 (1959) 432; 52, No. 2 (1960) 183; *Spec. Publ., Industr. Engng Chem., Amer. Chem. Soc.*, 6 pp., February 1960; *Industr. Engng Chem.* 53, No. 2 (1961) 166; *Spec. Publ., Industr. Engng Chem., Amer. chem. Soc.*, February 1961
- <sup>39</sup> WILLIAMS, T. J. *Systems Engineering for the Process Industries*. 1961. New York: McGraw-Hill
- <sup>40</sup> WILLIAMS, T. J. Ed., *Process Dynamics and Control, Chem. Engng Progr. Symp. Ser.* 57, No. 36 (1961) 172
- <sup>41</sup> WILLIAMS, T. J. Computers, automation, and process control. *I & EC Annual Review Supplement* (1962) 140; *Amer. chem. Soc.* (October 1962)
- <sup>42</sup> WILLIAMS, T. J. and MORRIS, H. J. A survey of the literature on heat exchanger dynamics and control. *Chem. Engng Progr. Symp. Ser.* 57, No. 36 (1961) 20
- <sup>43</sup> WILLIAMS, T. J. and OTTO, R. E. A generalized chemical processing model for the investigation of computer control. *Amer. Inst. elect. Engrs Spec. Publ., S 132, Computers in Control*, 130-145, September 1961
- <sup>44</sup> ZÁVORKA, J. The dynamic properties of rectification stations with plate columns. *Automatic and Remote Control* 1963. London; Butterworths: Munich; Oldenbourg

# Determination of System Dynamics by Use of Adjustable Models

E. BLANDHOL and J. G. BALCHEN

## Summary

A method for determining system dynamics by using a simulated model with adjustable parameters is described, and practical problems dealt with. Application of the statistical computer ISAC greatly simplifies the experimental work. The usefulness of the method is illustrated by some experimental results. A second-order system containing four parameters, including a pure time delay, was easily identified, with all parameter errors less than 3 per cent.

## Sommaire

Une méthode de détermination de la dynamique d'un système au moyen d'une simulation à l'aide d'un modèle à paramètres réglables, est décrite et certains aspects pratiques de l'application de cette méthode sont indiqués.

L'utilisation du calculateur statistique ISAC simplifie beaucoup le travail expérimental. L'utilité de cette méthode est mise en évidence par un certain nombre de résultats expérimentaux. Un système du 2<sup>e</sup> ordre, avec quatre paramètres dont un retard pur, a été facilement identifié, l'erreur de la détermination de chacun des paramètres étant inférieure à 3 pour cent.

## Zusammenfassung

Der Aufsatz beschreibt ein Verfahren zur Ermittlung des dynamischen Verhaltens von Systemen durch Verwendung eines nachgebildeten Modells mit einstellbaren Parametern und betrachtet praktische Probleme. Die Benutzung des statistischen Rechners ISAC vereinfacht den Untersuchungsaufwand beträchtlich. Einige Versuchsergebnisse zeigen den Nutzen der Methode. Bei einem System zweiter Ordnung, das vier Parameter und eine reine Totzeit enthält, konnten auf einfache Weise alle Parameter mit einem Fehler, der kleiner als 3% ist, bestimmt werden.

## Introduction

In recent years, the method of determining the dynamic properties of physical systems by using a simulated model with adjustable parameters has been given considerable attention. Much work has also been devoted to the development of more, or less, ingenious automatic adjustment procedures, so that the method could, for instance, be included in an adaptive control system.

Most of the rather few reported experiments have employed quite complicated and expensive equipment both for the model and the adjusting mechanism<sup>1-5</sup>. All that is needed in many cases, however, is some relatively simple equipment used in connection with a model which is manually adjusted to optimum. The method may then be a very powerful one for studying the dynamics of industrial processes and other non-adaptive systems. This paper shows that the statistical computer ISAC<sup>6, 7</sup> in connection with an analogue model, is well suited for this purpose. A more detailed account is found in the technical report<sup>8</sup>.

## Principle of the Model Method

The general principle of what, in this paper, is called the 'model method', is illustrated in *Figure 1*.

We want to estimate the dynamics of some physical system  $g$  by applying a test signal  $x(t)$  during some finite time  $T$ . The system output contains additive noise  $n(t)$ , so that only  $z(t) = y(t) + n(t)$  is measurable. Now a model  $h$  of the system is simulated, and the input  $x(t)$  is applied to the system and the model in parallel. The model output  $u(t)$  is then compared with the noisy system output  $z(t)$  according to some error criterion, and the model parameters are adjusted until the error becomes a minimum, following some pre-selected strategy. The resulting model configuration is said to be the 'optimum' approximation to the system.

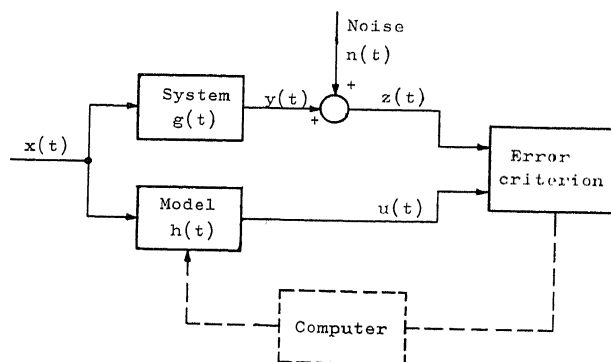


Figure 1. Principle of the model method

The parameter adjustments must be based on the measured error. In automatic model adjustment schemes an 'adjustment computer' is inserted as a feedback link, as indicated in *Figure 1*. In simpler cases the computer may conveniently be replaced by a human operator; this is the principle employed in our experiments.

## Practical Problems

Efficient use of the model method requires careful consideration of a number of practical problems, some of which are mentioned below.

### Choice of Model Structure

Perhaps the most important question when using the model method is: 'What should the model structure be?'

The process identification problem may basically be posed in two different ways:

(1) Having a physical system whose structure is known, but with unknown parameters. When these parameter values are required, the model structure should, of course, be chosen equal

to the system structure. If some simplification of the model is necessary, the kind of structural approximation made is always known.

(2) In other cases the system may be a 'black box' with a more or less unknown structure. In this event the system dynamics are merely approximated by some mathematical model of specified structure. A model transfer function with two poles, one zero, and a pure time delay may, for instance, be specified. A model satisfying these requirements may be synthesized in many ways and the resulting model structure may be quite different from the true system structure, even if the order of the model transfer function has been correctly chosen. On the other hand, it was experimentally observed that the type of model structure has a great influence on the accuracy and speed of convergence of the method. The choice of model structure should therefore be judged from this point of view. The model should also be chosen according to the kind of information desired. Thus, if the impulse response is required, a time domain model should be used, which may give an excellent approximation to  $g(t)$ , although its transfer function may be quite different from that of the system.

#### Time Scaling

Many systems, for instance industrial processes, have such large time constants that a model cannot be simulated on an analogue computer with reasonable component values. Moreover, experimental optimization of the model in real time would take a prohibitively long time. Some kind of time scaling is then necessary.

#### Error Criterion

A number of different error criteria have been proposed for use in model adjustments. Denoting the error by  $e(t)$ :

$$e(t) = z(t) - u(t) \quad (1)$$

the criterion  $\varepsilon$  used in our experiments was the integrated square error

$$\varepsilon = \int_0^T e^2(t) dt \quad (2)$$

which yields least-squares estimates for the model parameters<sup>9</sup>. Moreover, when the model output  $u(t)$  is a linear function of the model parameters, parabolic minima are obtained, which is of great practical importance.

#### Initial Conditions

When model adjustments are based on a finite record which is cyclically repeated, as in our experiments, the problem of what initial conditions to apply to the model in each run occurs. If possible, the measurements should be made with zero system initial conditions. Sometimes, for instance when measuring during normal system operation, this may be impossible. Then, if the initial conditions cannot be neglected or evaluated separately, they must be included in the model as additional adjustable parameters.

#### Adjustment Strategy

For  $n$  adjustable parameters the error criterion  $\varepsilon$  represents a surface in  $(n+1)$ -dimensional space, with an absolute minimum corresponding to the optimum set of parameter

values. The adjustment strategy should lead uniquely to this absolute minimum when starting from arbitrary initial values. Possible sources of difficulty are: (a) existence of local minima, (b) poor sensitivity of  $\varepsilon$  to some parameters, (c) poor convergence for certain model structures and strategies, and (d) lack of orthogonality, i.e. dependence of optimum value of one parameter on other parameter values.

When using an automatic 'adjustment computer' the strategy may involve quite complicated mathematical calculations. Examples of such strategies are the method of steepest descent<sup>3, 5</sup>, and the gradient method used in Feldbaum's optimizer<sup>2, 10</sup>.

For manual optimization another kind of strategy is desirable, avoiding mathematical calculations or simultaneous adjustment of several parameters. One obvious possibility, used with success in our experiments, and also suggested by Levin<sup>9</sup>, is a cyclical adjustment of the parameters, one at a time, until local minima occur. Using ISAC, the error measure  $\varepsilon$  is plotted on an  $xy$  recorder. The parameter adjustments are then based quite simply on visual inspection of the plotted  $\varepsilon$  points.

If the error  $e(t)$  is a linear function of the adjustable parameters  $\alpha_1, \alpha_2, \dots, \alpha_n$ , the error measure  $\varepsilon$  becomes, in matrix notation, a *quadratic form*. Then the cyclical adjustment strategy is equivalent to the Gauss-Seidel iteration procedure for solving the set of linear condition equations for a minimum in  $\varepsilon$ :

$$\frac{\partial \varepsilon}{\partial \alpha_i} = 0 \quad (i = 1, 2, \dots, n) \quad (3)$$

Since, from the definition (2),  $\varepsilon$  is positive definite, it has a unique parabolic minimum and the adjustment procedure is known to converge. When  $\varepsilon$  is a non-linear function of the parameters  $\alpha_i$  the equation set (3) will still be approximately linear in the vicinity of optimum.

When  $\varepsilon$  is a quadratic form it may, in two dimensions, be represented by a family of ellipses  $\varepsilon = \text{const}$ . The approach to optimum will then, with cyclical parameter adjustments, look as in Figure 2. The adjustment of one parameter will proceed until the trajectory becomes tangent to the family of ellipses. The loci of such points will be two diameters in the ellipses, as indicated in the figure. The speed of convergence clearly depends on the angle  $\phi$  between these diameters. In Figure 2 (a) the convergence is good, whereas in (b), where the performance surface has the form of a narrow 'valley', the convergence is very poor. The shape and orientation of the ellipses and the angle  $\phi$  may be expressed in terms of the elements in the coefficient matrix  $R$  for  $\varepsilon$ .<sup>8</sup> These matrix elements depend in a rather complicated manner on the input signal, the system function, the output noise and the integration time.

In Figure 2 the parameter axes do not coincide with the semi-axes of the ellipses. When they do, the parameter adjustments become orthogonal. In that case, the relative sensitivity of  $\varepsilon$  to the various model parameters is directly related to the corresponding eigenvalues in the coefficient matrix  $R$ . The optimum values of those model parameters corresponding to the smallest eigenvalues will be most sensitive to noise. The analytical requirement for orthogonal adjustments is that  $R$  be a diagonal matrix.

#### Use of the Statistical Computer ISAC in Model Experiments

ISAC<sup>6, 7</sup> is a self-contained analogue computer, primarily intended for automatic computation of correlation functions,

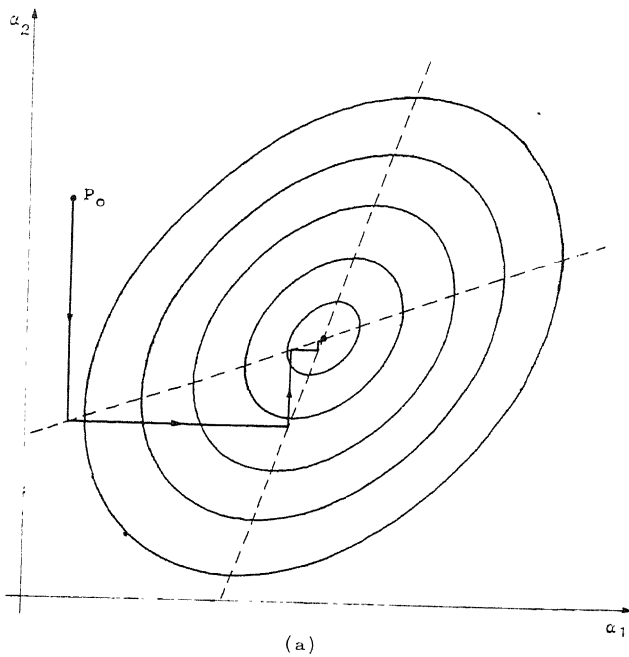
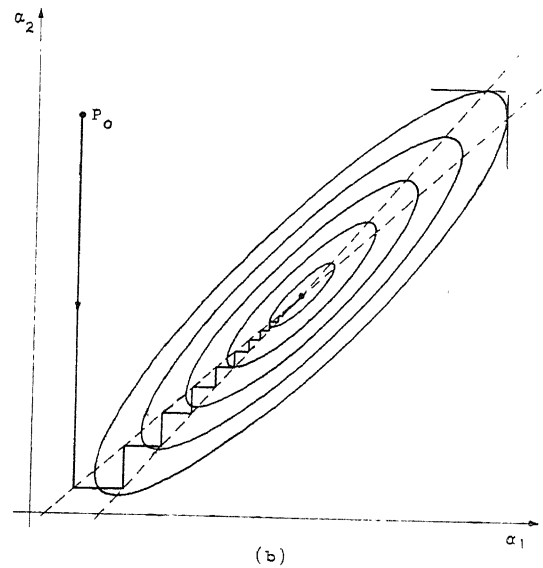


Figure 2. Approach to optimum with cyclical adjustment of parameters, for two different shapes of the  $\varepsilon$  surface



power spectra or amplitude distribution functions. It consists of a special-purpose magnetic tape recorder, electronic computing units and an  $xy$  recorder. One to three input signals are simultaneously recorded on a magnetic tape loop. There are eight recording tape speeds, covering a range of 1:1024. Playback is always performed at full tape speed. Maximum effective recording time is 68 min, playback revolution time 5 sec. Input frequency range is 0–200 c/sec at full tape speed, being proportional to tape speed. Two multi-track magnetic heads are used, one ( $A$ ) fixed, the other ( $B$ ) movable along a slide to obtain the time displacement  $\tau$  during correlation. The results are plotted on the  $xy$  recorder, and the automatic computation is controlled by relay circuits, triggered from two small holes in the tape loop\*.

The use of ISAC in model experiments is illustrated in Figure 3. The input  $x(t)$  and the output  $y(t)$  of the system to be studied are recorded on a tape loop, and then played back continuously during the adjustment procedure. The playback signal  $x_1$ , corresponding to the system input, is applied to the model, simulated on an analogue computer. The other playback signal  $y_1$  is subtracted from the model output to yield  $e(t)$ , which is then squared and integrated. By using the integrator in ISAC

the resulting value of  $\varepsilon$  may be automatically plotted, one point per revolution of the loop. Operation and resetting of the model is controlled by the triggering circuits in ISAC, so that identical initial conditions are maintained in all runs.

The time-scaling problem is easily solved by recording the test signals at a sufficiently low tape speed. The model is then simulated in playback time scale, where reasonable component values may be used.

The inclusion of a pure time delay in a model is usually a difficult problem. In ISAC this may easily be achieved by simply displacing head  $B$  during playback, yielding a delay  $0 \leq \tau \leq 320$  msec in playback time scale, or up to  $0 \leq \tau \leq 320$  sec in real time.

### Experimental Results

The applicability and versatility of the model method were studied by making quite an extensive series of experiments, covering several combinations of system structures, model structures, input signals, etc. For checking purposes all measurements were made on known systems, simulated on the analogue computer with zero initial conditions. One first-order and one second-order system were investigated, with or without a pure time delay included.

\* ISAC is now commercially available.

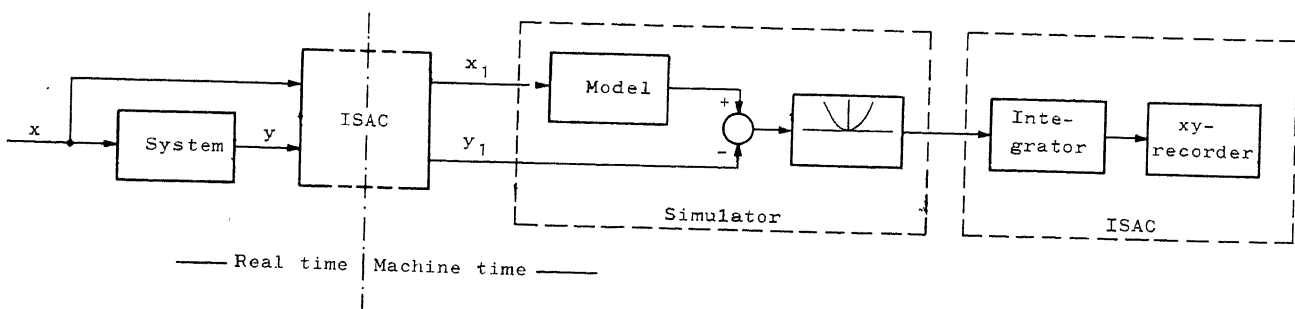


Figure 3. Use of ISAC in model experiments

An input signal  $x(t)$  was used with Gaussian noise of bandwidth 10 rad/sec. The playback-recording time-scaling factor  $\rho$  was 64, giving a recording time of 256 sec. The output noise  $n(t)$  was also Gaussian of bandwidth 10 rad/sec, statistically independent of the input  $x(t)$ . Recordings of the three signals  $x$ ,  $4y$  and  $4z$  are shown in *Figure 4*.

In the model transfer function (5)  $\beta_0$  or  $\beta_1$  must be normalized to 1. It was experimentally observed that in order to obtain useful results one must always use  $\beta_0 = 1$ . The reason is that for most input signals,  $\varepsilon$  will be more sensitive to errors in gain than in the time constant. Thus, when the gain depends on two parameters,  $\alpha_0$  and  $\beta_0$ , and adjustments are started from arbitrary initial values, one will, after only one or two adjustment cycles, end up with a ratio  $\alpha_0/\beta_0$  close to the theoretical gain, and a fairly low value of  $\varepsilon$ . But then no further reduction in  $\varepsilon$  can be observed, although the absolute values of  $\alpha_0$ ,  $\beta_0$  and  $\beta_1$  may all be completely wrong. This also applies to higher-order models.

The playback circuit diagram is shown in *Figure 5*. The d.c. values of the two ISAC demodulator outputs are suppressed, and some extra amplifiers are added to obtain the necessary amplification and uniform loading of all parameter potentiometers.

For the signals in *Figure 4* the model was optimized with and without output noise. Starting with the arbitrary initial values  $\alpha_0 = 0.6$ ,  $\beta_1 = 0.714$ ,  $\tau = 20$  mm, the following optimum values with noise were found after a total of 10 parameter

Referred to real (recording) time scale the system transfer function  $G_1(s)$  was chosen as

$$G_1(s) = \frac{a_0}{b_0 + b_1 s} e^{-\tau_0 s} = \frac{1}{1+s} e^{-6.11s} \quad (4)$$

where  $\tau_0 = 6.11$  sec corresponds to 30 mm displacement of head  $B$ . The corresponding model transfer function  $H_1(s)$  is written

$$H_1(s) = \frac{\alpha_0}{\beta_0 + \beta_1 s} e^{-\tau s} \quad (5)$$

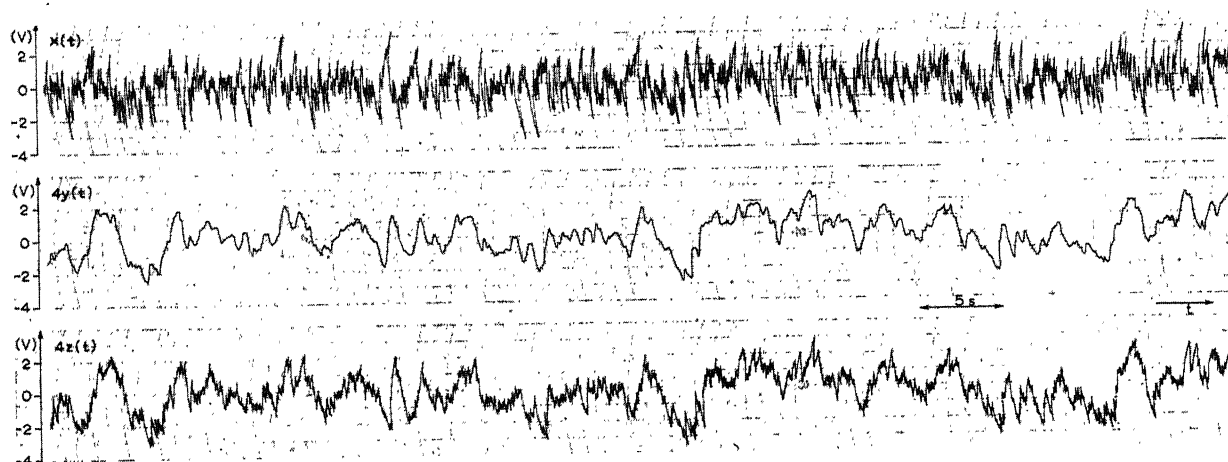


Figure 4. Test signals for first-order system

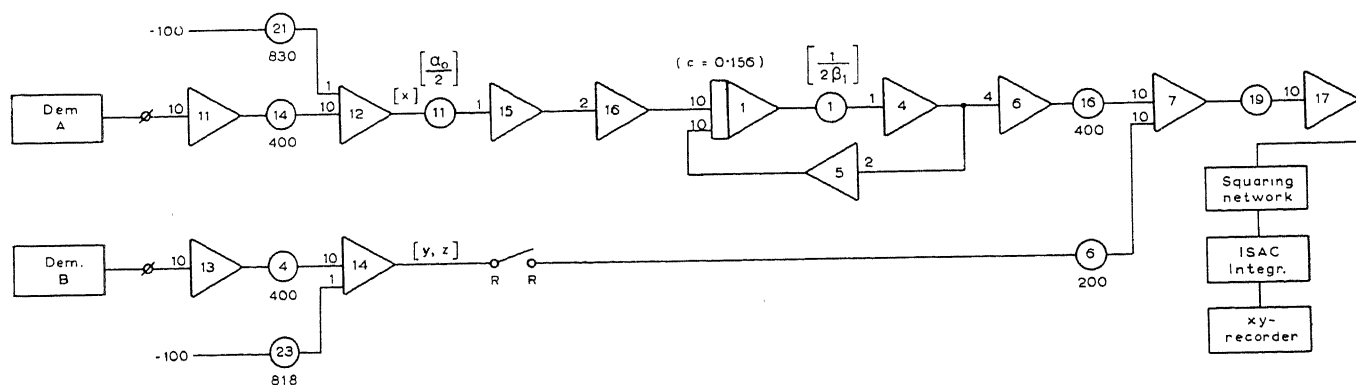
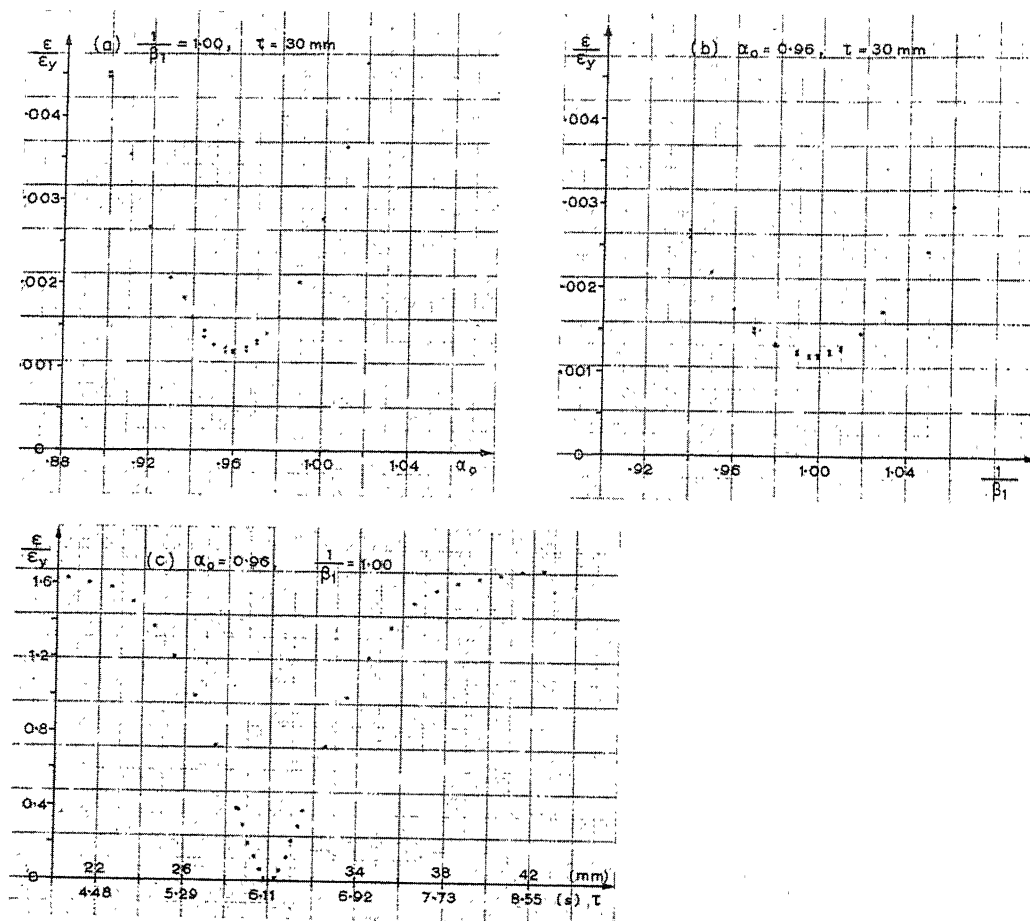


Figure 5. Circuit diagram for first-order transfer function model

Figure 6. Optimum  $\varepsilon$  curves for first-order model without noise

adjustments (finding four local minima for  $\alpha_0$  and  $\beta_1$ , and two local minima for  $\tau$ ):

$$\alpha_0 = 0.980 \quad \beta_1 = 1.020 \quad \tau = 30.0 \text{ mm} \quad R = 0.081 \quad (6)$$

Here  $R$  is a figure of merit, defined as

$$R = \frac{\Delta \varepsilon_{\min}}{\varepsilon_z} = \frac{\int_0^T \varepsilon_{\min}^2(t) dt}{\int_0^T z^2(t) dt} \quad (7)$$

where  $\varepsilon_{\min}$  is the minimum value of  $\varepsilon$  obtained with the optimum model, and  $\varepsilon_z$  is the value of  $\varepsilon$  when the model is identically zero. Note that  $\varepsilon_z$  is easily obtained experimentally by simply breaking the connection from pot. (16) to amplifier (7) in Figure 5.

When the adjustments were started with  $\alpha_0$  or  $\beta_1$  the completely 'wild' values  $\alpha_0 = 0$  or  $\beta_1 = 4.17$  were obtained in the first run, with only a slight decrease in  $\varepsilon$ .  $\tau$  at once converged to the value 30.25 mm, decreasing  $\varepsilon$  by a factor of 8, in spite of the initial errors in  $\alpha_0$  and  $\beta_1$ . The reason for this is that the initial error in  $\tau$  is so large that the outputs  $y$  and  $u$  are practically uncorrelated, no matter what values we give  $\alpha_0$  and  $\beta_1$ . The important conclusion is that when studying a system where time delays may occur, the model should be tested for  $\tau$  at an early stage of the optimization process.

Starting with the same initial values without noise, 10 parameter adjustments lead to the optimum values:

$$\alpha_0 = 0.960 \quad \beta_1 = 0.994 \quad \tau = 30.0 \text{ mm} \quad R = 0.0012 \quad (8)$$

The plotted optimum  $\varepsilon$  curves are shown in Figure 6. The curves for  $\alpha_0$  and  $1/\beta_1$  are seen to be approximately parabolic. The  $\tau$  curve (note different ordinate scale) becomes flat for large errors in  $\tau$ , when  $y$  and  $u$  become uncorrelated, as mentioned above.

To obtain the high sensitivity used in the curves for  $\alpha_0$  and  $\beta_1$  extreme accuracy is needed in the experimental equipment. All  $\varepsilon$  points in the three curves were plotted three times, but the spread is hardly noticeable. Smooth curves may also easily be drawn through all computed points. Both checks verify the high accuracy and stability existent in the computer ISAC.

The values found in (6) and (8) are all within 4 per cent of their theoretical values. The largest error occurs in  $\alpha_0$ , but further tests suggested the presence of a systematic error, always giving  $\alpha_0 < 1$ .

A recording time of 256 sec is quite large in relation to the system time constant and the input noise bandwidth. In this case evaluation of the model parameters from statistical correlation functions, also computed on ISAC, gave about 3 per cent error in gain and 16 per cent error in time constant<sup>8</sup>.

To test the accuracy of the model method for short recording times, when statistical techniques are quite impossible, a recording time of only 8 sec was tried, with all other test conditions unchanged, except that no time delay was used. The following quite satisfactory results were obtained, with rapid convergence:

	$\alpha_0$	$\beta_1$	$R$	
Without noise	0.984	1.025	0.0007	(9)
With noise	1.110	1.099	0.065	

### Second-order System

In recording time scale,

$$G_2(s) = \frac{a_0}{b_0 + b_1 s + b_2 s^2} e^{-\tau_0 s} = \frac{1}{1 + 1.4s + 0.4s^2} e^{-\tau_0 s} = \frac{1}{(1+s)(1+0.4s)} e^{-\tau_0 s} \quad (10)$$

where  $\tau_0 \approx 30.5$  mm. The same structure was used in the model, with  $\beta_0$  normalized to 1. The appropriate modification of the circuit diagram in *Figure 5* is shown in *Figure 7*;  $x(t)$  and  $n(t)$  were maintained as for the first-order system.

First a 256 sec recording was studied. As arbitrary initial parameter values were chosen  $\alpha_0 = 0.6$ ,  $\beta_2 = 0.286$ ,  $\beta_1 = 2.4$ ,  $\tau = 20$  mm. The convergence was rapid, and without noise the following values were obtained after 18 parameter adjustments:

$$\begin{array}{lll} \alpha_0 = 1.002 & \beta_1 = 1.440 & R = 0.0008 \\ \beta_2 = 0.394 & \tau = 30.55 & \end{array} \quad (11)$$

As for the first-order system  $y$  and  $u$  were found to be uncorrelated for  $\tau = 20$  mm.  $\tau$  thus had to be adjusted first, at once

leading to  $\tau = 30.75$  mm. The plotted optimum  $\varepsilon$  curves are shown in *Figure 8*. Due to the high model sensitivity a slight spread in the triple-plotted points is observed.

The convergence was satisfactory with noise also, yielding the final result:

$$\begin{array}{lll} \alpha_0=1.014 & \beta_1=1.440 & R=0.09 \\ \beta_2=0.417 & \tau=30.55 & \end{array} \quad (12)$$

The optimum parameter values found in (11) and (12) are seen to be very accurate, the maximum error being 2.9 per cent without noise and 4.3 per cent with noise.

When trying to evaluate the second-order system by correlation techniques, the cross correlation function must be approximated by three exponential terms<sup>8</sup>. Due to the highly non-orthogonal nature of the exponential functions this would require an excessive data accuracy which by no means is attainable in experimental correlation functions, and was thus not attempted.

Finally, an 8 sec recording was studied, with no time delay in the system. Even then the convergence and accuracy were quite satisfactory both with and without noise. The following results were obtained:

	$\alpha_0$	$\beta_1$	$\beta_2$	$R$
Without noise	0.962	1.400	0.396	0.002
With noise	0.940	1.400	0.357	0.34

(13)

### Parameter Determination in Multi-input and Non-linear Systems

In the previous sections the noise was regarded as being added to the system output and being non-measurable. Actually, in many practical cases, what has here been considered as noise

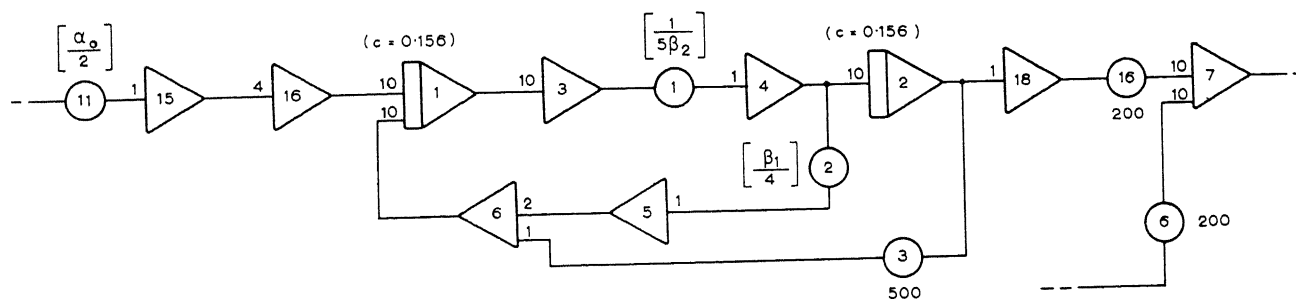


Figure 7. Modification of Figure 5 for second-order transfer function model

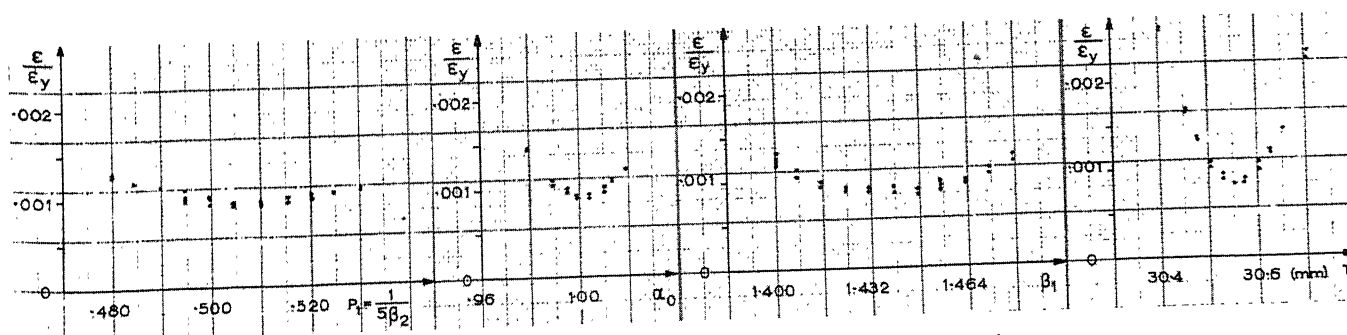


Figure 8. Optimum  $\varepsilon$  curves for second-order model without noise

may well be the response of another measurable input to the system. In such cases it is certainly advantageous to record this input ( $x_2$ ) as well as the ordinary input ( $x_1$ ) considered before. Now a model having two channels (I and II, from  $x_1$  and  $x_2$  respectively) can be made. The model adjustment procedure could be as follows: first the channel I from  $x_1$  to  $u$  is adjusted to optimum, regarding  $x_2$  as noise. This channel is then maintained while adjusting channel II till optimum is reached. Maintaining channel II, a new adjustment of channel I can be made leading to an almost complete compensation of the effect of the 'noise' input  $x_2$ . The computer ISAC, having three recording channels, is well suited for such purposes as well.

Identification of non-linear systems has received very little attention in the literature so far. Model adjustment procedures as outlined here appear to be quite promising in this respect as well. If the system structure is known, say by its differential equation, and some non-linearity appears in it which can be characterized by a small number of parameters, then the problem of adjusting a non-linear model is essentially the same as in the linear case. Certainly the magnitude of the signals now becomes more important than in the linear case. The success in characterizing a non-linear element is thus dependent upon having signals that actually reveal the non-linear phenomena.

Even though only introductory investigations have been made so far on these latter problems quite interesting results seem to be within reach.

## Conclusions

It has been demonstrated that the computer ISAC is extremely well suited for model experiments, providing recording and cyclical playback of the test signals, with a time scaling factor up to 1,024, and easily permitting the inclusion of a time delay in the model. The results obtained for the practically important second-order model, containing four adjustable parameters, including a pure time delay, give a very promising illustration of the potentialities of the model method, when precision equipment and correct experimental techniques are used.

## DISCUSSION

A. HAKALA, *Kymmene Aktiebolag, Kuusankoski, Finland*

This method and device could be a useful tool for determining the dynamics of processes during their normal operation.

Have the authors made any actual tests in the field, and if so what is their idea about the capabilities of this method and equipment?

K. S. P. KUMAR, *C.I.S.L. School of Electrical Engineering, Purdue University, Lafayette, Indiana, U.S.A.*

(1) The modelling methods can be viewed as multi-point boundary value problems and very effective computational results obtained via the method of quasi-linearization<sup>1</sup>. Suppose the control system to be given by

$$\dot{x} = f(x, u, p, t) \quad (1)$$

where  $u$  is the control vector,  $x$  is the state vector and  $p$  are some unknown constant parameters of the system like, for example, time constants, gains, etc. As  $p$  are constants, they are written as

$$\dot{p} = 0 \quad (2)$$

*The material presented here is based upon the results of a research project at the Division of Automatic Control, the Technical University of Norway, sponsored by the Royal Norwegian Council for Scientific and Industrial Research. The support is highly acknowledged.*

## References

- GOODMAN, T. P., and RESWICK, J. B. Determination of system characteristics from normal operating records. *Trans. Amer. Soc. Mech. elec. Engrs.* 78 (1956), 259-71
- STAKHOVSKII, R. I. Twin-channel automatic optimizer. *Automat. Rem. Contr.* 19, No. 8 (Aug. 1958), 729-40
- MARGOLIS, M., and LEONDES, C. T. On the theory of adaptive control systems; the learning model approach. *Automatic and Remote Control*. Vol. II, pp. 556-63. 1961. London; Butterworths
- GIBSON, J. E. Self-optimizing or adaptive control systems. *Automatic and Remote Control*. Vol. II, pp. 586-95. 1961. London; Butterworths
- TAYLOR, W. K. An experimental control system with continuous automatic optimization. *Automatic and Remote Control*. Vol. II, pp. 686-93. 1961. London; Butterworths
- BLANDHOL, E., HESTVIK, O., and MOHUS, I. *A Description of the Statistical Computer 'ISAC'*. Automatic Control Laboratory, Norwegian Institute of Technology, Trondheim, 1 (Nov. 1959), 2 (Dec. 1960)
- BALCHEN, J. G., and BLANDHOL, E. On the experimental determination of statistical properties of signals and disturbances in automatic control systems. *Automatic and Remote Control*. Vol. II, pp. 788-96. 1961. London; Butterworths
- BLANDHOL, E. On the use of adjustable models for determination of system dynamics. *Tech. Rep. No. 62-5-D. Div. Automat. Contr.*, Technical University of Norway, Trondheim (March 1962)
- LEVIN, M. J. Estimation of the characteristics of linear systems in the presence of noise. *Dr. Sc. Engng. Dissert.*, Dept. elect. Engng., Columbia University, New York, N. Y. (July 1959)
- FELDBAUM, A. A. Automatic optimizer. *Automat. Rem. Contr.* 19 No. 8 (Aug. 1958), 718-28

Measurements are made on the control vector and state vector. It is unnecessary to have access to the entire state vector. It is enough if only a few components of it, or linear combinations thereof, are measured. Now eqns (1) and (2) are solved with the given boundary conditions on  $x$ . The method of quasi-linearization remains exactly the same even if eqn (1) is non-linear.

Both cases, known structure and unknown parameters and unknown structure and unknown parameters, have been examined by the above method at the Control and Information Systems Laboratory at Purdue University with excellent results. These are under preparation for publication. Evidence shows that this method has great promise towards on-line control.

(2) Did the authors try any steepest descent methods and if so how did they work out?

## Reference

- KALABA, R. On non-linear differential equations, the maximum operator, and monotone convergence. *J. Math. Mech.* 8 (1959), 519-574



A. J. NIEMI, *University of Oulu, Hoikantie 14 A 7, Oulu, Finland*

The impulse response of a system is simply the inverse transform of its transfer function and hence contains the same parameters. With respect to an impulse input, the frequency domain model, i.e. the transfer function, and the time domain model are therefore essentially equivalent.

However, the authors recommend the use of a time domain model when looking for the impulse response of a system. I would like to know the reason for this preference and the real advantages gained by using the time domain model instead of a transfer function in this case.

H. GORECKI, *Academy of Mining and Metallurgy, Gramatyka 7/5, Krakow, Poland*

What is the effect of multiplicative noise in the input signal, when due to changes of system parameters, for instance?

Has there been any research into the identification of the system structure? For example, at the Chair of Automatic Control in the Academy of Mining and Metallurgy in Krakow we determine the structure of non-linear control systems or processes by the frequency response method. Suppose it is required to establish the equation of the plant:

$$\ddot{x} + a_2 \dot{x} + f(x) \dot{x} + a_0 x = 0 \quad (1)$$

or

$$\ddot{x} + a_2 \dot{x} + a_1 \dot{x} + g(x) = 0 \quad (2)$$

For these equations the frequency characteristics are determined as families of curves for different input signal amplitudes. In the case of eqn (1) the distance between the curves of the real part is proportional to this amplitude, whereas in the case of eqn (2) this is true for the curves of the imaginary part. The method might be extended to higher-order systems.

R. W. H. SARGENT, *Chemical Engineering Dept., Imperial College, London S.W. 7*

May I ask if the authors have made any experiments to test the convergence of their proposed method of alternate adjustment for multi-input systems?

It is easy to construct response surfaces for which the cyclical adjustment of variables will not converge on the optimum (even a local one), and one would intuitively expect the same type of behaviour in this multi-input problem.

R. J. A. PAUL, *College of Aeronautics, Cranfield, Bletchley, Bucks, England*

(1) First- and second-order systems: Since the parameter adjustments in the model are not orthogonal, how many iterations were necessary to achieve the optimum parameters? The necessity for finding the value of the transportation lag, as a first requirement, appears to be valid.

(2) Higher-order systems: Have the authors carried out any experimental investigations with high-order systems approximated with the simplified model comprising a transportation lag, a zero and a pair of conjugate complex poles? If so, were any indications evident regarding the limitations of this approach?

Investigations<sup>1</sup> indicate that if the poles of the high-order system are close together then it is extremely difficult to obtain a reasonable approximation with a simplified model.

## Reference

- <sup>1</sup> PAUL, R. J. A. Determination of dynamical models for adaptive control systems. College of Aeronautics, Note No. 128 (1962)

C. H. P. BROOKES, *Department of Engineering Science, Oxford University, Oxford, England*

At Oxford some work has been done using methods similar to those of Messrs. Blandhol and Balchen on models which are manually adjusted and optimized. The approach has been to investigate the accuracy required in a model in order for it to be able to predict either system performance or an optimum control setting. No attempt is made to discover the position of the system's poles or its gain; it is considered sufficient that the model is able to represent the system adequately.

Type 1 systems with unity feedback have been considered up to date with a simple lag compensation network of form  $(As + 1)/(Bs + 1)$  included in the forward path. The model was set up in the same way with similar compensation. It consisted of a Type 1 servo with variable gain and pole positions.

Step inputs were applied to both system and model and the model parameters were adjusted to minimize a function of the system-model error, usually the *ITAE* or *ISE*. Systems of up to tenth order, with and without complex open-loop poles, were simulated on an analogue computer and for each system several models of different complexity were used. Models were judged on their ability to predict the compensation which minimized the system *ITAE* and their prediction of the system output trajectory after a change in the compensation networks.

One interesting conclusion from this work is that if the model consists of poles confined to the real axis of the *s* plane then the best configuration for a Type 1 model used to predict higher-order systems has a pole at  $s = 0$  and a multiple pole somewhere along the real axis. This reduces the number of parameters to be adjusted in the modelling process to two, the gain and multiple pole position. The problem of modelling is eased considerably, especially when it is necessary to use models above third order.

The result applied for all the systems considered so far. The two variables may be orthogonalized easily by using standard techniques, but it has been observed experimentally that if the adjustments are made on the velocity constant and multiple pole position the settings are nearly independent. Model performance, of course, deteriorates as the difference in order between system and model grows, but for models above fourth order the deterioration is much less marked. The use of a second-order model in a set-up as described above yields very poor prediction even for fourth-order systems.

Random input signals and more complicated compensation networks have been tried and the results regarding model configuration are the same. It is envisaged that a model of a very complex system will consist of several of these multi-pole models linked together.

H. J. MEYERHOFF, *Pembroke College, Cambridge University, Cambridge, England*

Manual adjustment of the model parameters  $\alpha_i$  requires that their setting be non-interacting. For a model whose output is a linear function of the variable parameters, i.e.

$$\mu(t) = \sum_{i=1}^N \alpha_i \phi_i(t)$$

a simple self-adjusting procedure has been developed<sup>1</sup> which does not require that the setting of  $\alpha_i$  be non-interacting. Instead of determining the minimum value of

$$\frac{1}{T} \int_0^T e^2(t) dt$$

by successive adjustments to the parameters  $\alpha_i$ , the proposed self-adjusting procedure solves the simultaneous equations:

$$\frac{1}{T} \int_0^T e(t) \phi_i(t) dt \doteq 0 \quad i = 1, 2, \dots, N$$

by continuously adjusting the  $\alpha_i$ . The setting time for the identification procedure may be shortened at the expense of accuracy of the para-

meter settings, which corresponds to using shorter interval lengths of the input  $x(t)$ . A repetitive test record of the system input is not required, and the identification may be carried out continuously using the normal system input.

Where the model output is not a linear function of the model parameters, as is given in the examples with variable poles in the model transfer function

$$H_1(s) = \frac{\alpha_0}{\beta_0 + \beta_1 s} \cdot e^{-\tau s}$$

the equations for setting the model parameters automatically may be unstable outside the small lines region near the stable equilibrium point. There may also exist unstable equilibrium points.

## Reference

- <sup>1</sup> MEYERHOFF, H. J. Self adjusting process models. *Thesis*, Cambridge University (1963)

J. RAKOWSKI, *Institute of Power, Woloska 88 m 53, Warsaw 12, Poland*

The examples in the paper show how the model method can be used to determine the dynamics of one-input and one-output systems.

The following question therefore arises: how could this method be used to determine the dynamic properties of a multi-input and multi-output system. For example, in the case of a steam generator one has to consider at least three main input and three output signals, forming three main control loops, i.e. three separate controllers. The structure of this system is complicated. There could be also some difficulty with opening (disconnecting) more than one of the control loops during the period of the test.

Can the described model method be applied to determine the dynamics of such a system? How should one attack this problem?

K. REINISCH, *Institut für Regelungstechnik, Heideflügel 3, Dresden, Germany*

Applying deterministic signals (step functions and similar types), instead of stochastic ones, to the parallel compensation method investigated by the authors, one may use another adjustment strategy<sup>1</sup>. This gives the time constants (poles and zeros) of the plant by the first approximation step with an accuracy of about 5–10 per cent and therefore shortens the identification process. For a class of plants this method allows the structure of the plant transfer function to be identified. It would be interesting to know whether the authors did some work in this direction, too.

## Reference

- <sup>1</sup> REINISCH, K. Verwendung eines Modellregelkreises zur Gewinnung einfacher Bemessungsregeln für Regelkreise und zur Ermittlung der Kennwerte von Regelstrecken. *Zmsr* 5, H. 6 (1962), 245–251

A. A. FELDBAUM, *Institute of Automatics and Telemechanics, Kalanchevskaja 15a, Moscow, U.S.S.R.*

The paper is concerned with an extremely topical question and clearly demonstrates the convenience of determining the dynamic characteristics of plants with the aid of the ISAC computer. As regards the content of the paper, I make the following comments of a basically theoretical and also practical nature.

(1) What is the advantage of using an adaptive analogue for determining the dynamic characteristics of plants, as compared with other known methods? For idealized problems it is possible to find the optimal algorithm for processing the information coming from the plant input and output; this information is mixed with random noise.

It can be shown that the structural scheme which corresponds with the optimal algorithm does not match that examined in the paper. Hence the scheme of the adaptive analogue is not, from a theoretical viewpoint, the best one. Of course, building the scheme in *Figure 1* of the paper does not require the solution of any complex theoretical problems and this is its practical advantage.

But does not this scheme have any fundamental advantages? It would be interesting to know the authors' view of this.

(2) The paper examines a process of optimization which is effected manually. However, experience in the Soviet Union indicates that man copes badly with the complex problems of optimization (when there is a comparatively large number of variables, when the extremized function has complex form, when there is a large number of trials with various initial values, different variants of analogue structure, etc.). The automatic optimizers used in the Soviet Union in equipment for automatic synthesis of optimal systems are comparatively simple instruments, but the problems solved by this equipment are not within the power of man, who would attempt to solve them manually.

When using the method of two analogues connected in parallel, in which the parameter values differ by one trial step (as in Dr. Mesch's paper), an even simpler instrument, of the controller type, can be used.

P. M. E. M. van der GRINTEN, *Central Laboratory, Staatsmijnen, Geleen, Netherlands*

In your paper, as in many others on model updating, not much attention is given to the requirements imposed on the test signal  $x(t)$ . Yet the influence of the properties of the test signals (power spectrum and possibly cross-power spectrum), on the speed and accuracy of convergence, is presumably as equally important as the choice of the model structure. Your statement that an apparently good approximation of the impulse response sometimes corresponds to a poor approximation of the transfer function, will perhaps also find its explanation in the application of test signals of a rather narrow bandwidth.

H. A. BARKER, *Glasgow University, Glasgow W. 2, Scotland*

The method of model optimization described here appears to offer no advantages over methods in which the integral error-squared signal is used to adjust the model parameters directly by means of a simple feedback loop. These methods are applicable to linear and non-linear multivariable systems, and may be used for manual or automatic adjustment when the effects of model parameter adjustments have been made independent by an orthogonalization procedure, or for simultaneous automatic adjustment when these effects are not necessarily independent. The optimization time is less than that required for the models described in this paper, and it is not necessarily increased if the number of model parameters is increased. A single run of the system signal records is sufficient for the optimization to be accomplished, using any convenient time scale, or alternatively the signals which occur during system operation may be used directly.

## Bibliography

- LUBBOCK, J. K. and BARKER, H. A. A solution of the identification problem. *Proc. JACC*, Minneapolis (June 1963)
- BARKER, H. A. The use of orthogonal functions for the solution of optimization problems. *Ph. D. Thesis*, Cambridge University (1963)

F. MESCH, *Institut für Regelungstechnik, Darmstadt, Germany*

If the mean square value of the noise  $n(t)$  is large compared to that of the input signal  $x(t)$ , the proposed configuration of *Figure A* might fail completely. I want to suggest another configuration that eliminates the effect of the noise and, in addition, replaces the extremum-seeking (even function of parameter deviation) by a straightforward adjustment to zero (odd function of parameter deviation).

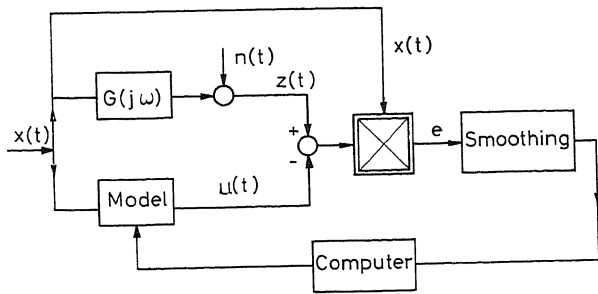


Figure A

Compared with Figure 3 of the paper, the squaring of the difference is replaced here by the multiplication of the difference with the input signal. It may be shown that to a first-order approximation

$$\overline{e(t)} = \overline{x(t) \cdot [z(t) - u(t)]} \approx \Delta K \cdot \int_{-\infty}^{\infty} S_{xx}(\omega) \cdot \frac{\partial G(j\omega; K)}{\partial K} d\omega$$

where  $K$  denotes the parameter to be adjusted, and where the parameter deviation  $\Delta K$  is assumed to be small. The integral forms a constant and expresses the sensitivity to the parameter  $K$ . The first factor  $\Delta K$  is proportional in sign and amplitude to the parameter deviation, so the model is simply adjusted for  $\Delta K = 0$ . This procedure, I feel, might be readily extended to more parameters.

B. P. TH. VELTMAN, *Technological University of Delft, Lorentzweg 1, Delft, Netherlands*

This clear and convincing presentation of the 'model method' seems to outperform the 'statistical method'. The authors' opinion about this point is explicitly mentioned in their paper. If by 'statistical method' the conventional correlation procedure is meant, this statement needs some more attention.

The measurement of a correlation function can also be seen as a model method: the model being a pure time delay, which is certainly a reasonable model if there is complete ignorance about the systems dynamics. The output signal of the system and of the time-delay model may be highly non-orthogonal, thus necessitating long observation times. However, if some dynamic properties of the system are known *a priori* one may use this information for prefiltering the correlator input signals to make them more orthogonal. Such a scheme is described briefly in the paper by A. van den Bos and myself where small bandwidth signals are used as correlator inputs. The parameter estimation time is in that case of the same order of magnitude as with the model method described by the authors. What I would like to stress is that the convergence time in parameter estimation is dominated by the *a priori* knowledge of the system and that it furthermore only depends on the available equipment which measuring strategy should be followed.

L. R. YOUNG, *Massachusetts Institute of Technology, Cambridge 39, Massachusetts, U.S.A.*

I would like to call the authors' attention to the very efficient method for parameter identification at time-varying processes by the multiple regression technique, presented by Dr. J. I. Elkind at the Joint Automatic Control Conference in New York, June 1962.

E. BLANDHOL and J. G. BALCHEN, *in reply*

We would like to express our most sincere thanks for the great interest taken in our paper by all the contributors to the discussion.

The subjects treated in the discussion may roughly be divided into five groups. We have therefore found it convenient to comment on each group separately.

However, we start with a brief summary of some experimental results which were included in the presentation of the paper at the Congress:

Quite recently some preliminary results were obtained from a model study of a non-linear industrial process, shown in principle in Figure A. A liquid flows through a tube between a pressurized tank and an evaporator. The flow is turbulent, giving a quadratic pressure loss due to friction. The driving force is the tank pressure minus the vapour pressure in the evaporator. The differential equation of this system is shown below Figure A. The unknown parameters are  $m$  and  $\xi$ ;  $m$  can be estimated fairly well, but the value of  $\xi$  is rather uncertain.

The process oscillates, with a period of about 3 min. These oscillations were used as test signals in a model to determine  $m$  and  $\xi$ .

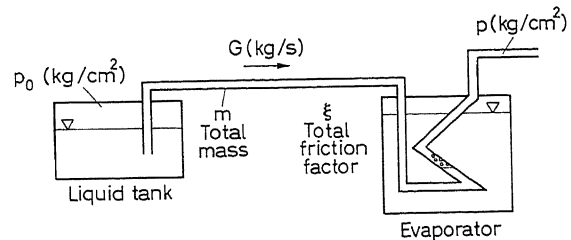


Figure A. Non-linear industrial process

Differential equation:

$$\frac{dG}{dt} = \frac{1}{m} [9.3(p_0 - p) - 0.051 \xi G^2]$$

Time scaling factor:	$\beta = 115$	
Model parameters:	$\frac{1}{m}$	$\xi$
Expected values:	0.0264	122 (uncertain)
Optimum values obtained:	0.0274	200
Per cent error:	3.8	65
Normalized error measure	$R = \frac{\varepsilon_{\min.}}{\varepsilon_x}$	
at optimum:	$R = 0.025$	

The time scaling factor  $\beta$ , the expected values of the model parameters and the values actually obtained are all listed below Figure A. It is seen that the optimum value of  $1/m$  is quite close to the expected one, whereas  $\xi$  turned out to be considerably higher than assumed. But the normalized error  $R$  at optimum is only 2.5 per cent, indicating that the model approximation obtained is quite satisfactory.

Figure B shows the various signals recorded from the system and the model, and Figure C shows one of the minimum curves obtained. The convergence seems satisfactory in the vicinity of optimum. A more general study of the convergence over a broader parameter range, or of other model configurations, has not been made yet, but these questions will be further investigated.

Our conclusion to these last experiments is that the model method seems to work well also for non-linear systems, probably provided certain assumptions regarding signals, structures and convergence are satisfied. This is of great practical importance, since most processes and systems to be studied will probably contain some non-linearities.

The above remarks also provide an answer to the question from Mr. Hakala. Now we proceed to the different groups of comments.

#### Input and Noise Characteristics

We certainly agree with Dr. van der Grinten about the sensitivity of the model method to input signal characteristics, as also mentioned in the paper when discussing the coefficient matrix  $R$  for  $\varepsilon$ . But one of

the major advantages of the model method, compared with, for instance, statistical methods or specific transient tests, is, in our opinion, that it usually works well with quite short samples within a large class of test signals, including the normal process operating signals.

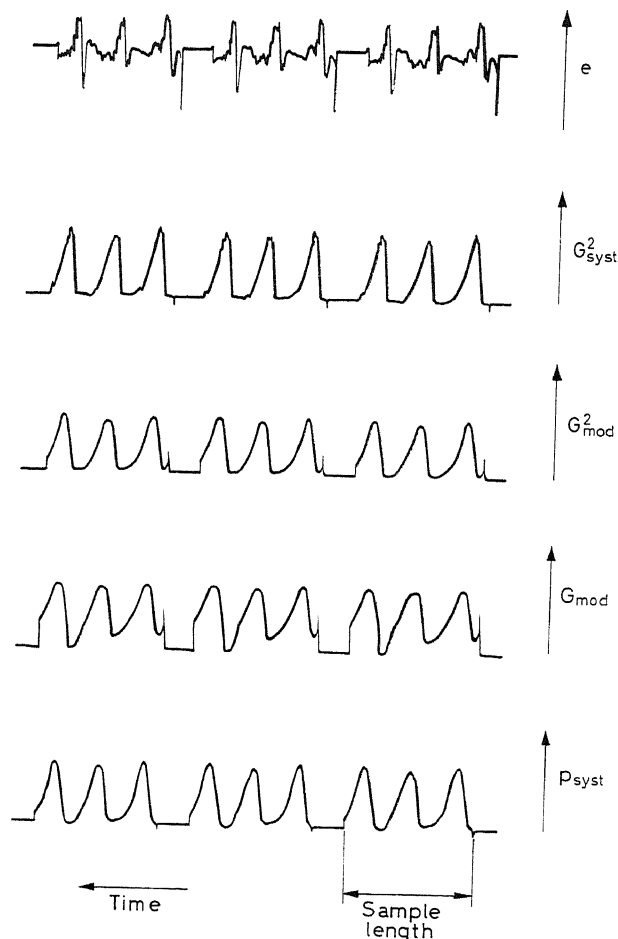


Figure B. Recorded signals from system and model

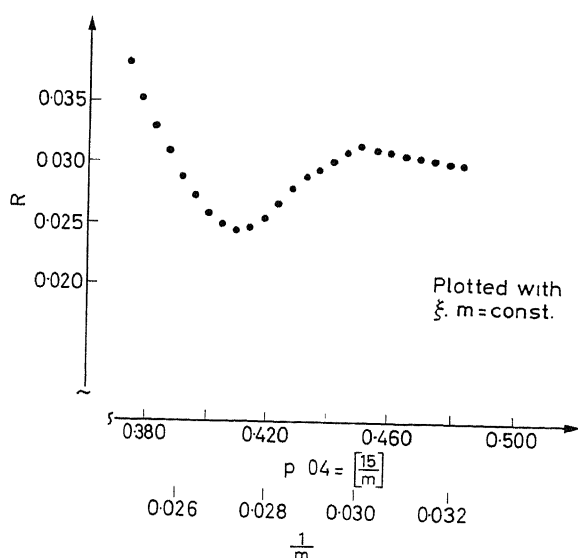


Figure C. Experimentally obtained minimum curve

Our method, and probably any model method, assumes constant model parameters throughout each run of the test signal. Thus, if multiplicative noise due to changes in system parameters is present, as mentioned by Professor Gorecki, this is equivalent to having a time-variable system. The resulting model will then be a mean-square approximation to the variable system over the test period.

#### System Structure vs. Model Structure

The main problems here seem to be:

(1) With model structure equal to a known system structure, how complicated can that structure be before the convergence deteriorates? We believe that not only the order of the system, but also the number of adjustable parameters are of importance, since some parameters may be known *a priori*. In general, poorer convergence must be expected for higher-order systems, containing more parameters, but the upper practical limit should be 4–5 adjustable parameters.

(2) What is the effect of using a simplified model for a complicated, known system structure? It seems impossible to give a general answer to this very important problem, although certain conclusions may be drawn from theory or from performed experiments. The results mentioned by Mr. Paul could thus be readily expected. We find the work reported by Mr. Brookes interesting, since any possible reduction in the number of model parameters is important. The authors have not yet made any experiments on this problem.

(3) What model structure should be used when the system is a 'black box'? This is the problem of designing a mathematical model, as stressed in the paper. One possibility is to try several different model structures, and select the one giving the best approximation. In this way the model method may be used also for optimizing model structure, not only the parameters. This would also be the model method equivalent to Professor Gorecki's frequency method for choosing between his eqns (1) or (2).

(4) What is the difference between a time-domain and a frequency-domain model? Here some confusion seems to exist. Theoretically, the impulse response and the transfer function are essentially equivalent, so that either model might be used. However, in practice we are dealing with parameter estimates only, calculated for finite measuring time and in the presence of noise. We always want to find optimum estimates, but a method which is optimum in one domain is not likely to be optimum in the other. This can be intuitively understood by realizing that the frequency-domain model parameters are complicated functions of the time-domain parameters, and vice versa. Hence it is very difficult to calculate the propagation of parameter uncertainties from one domain into the other. Moreover, parameter adjustments which may be nearly independent (orthogonal) in one domain will certainly not be so in the other domain. We hope this provides an adequate answer to Messrs. Niemi and van der Grinten.

#### Adjustment Strategy and Convergence

Also on this point there seem to be some misunderstandings. In its present form our model method is not a self-adjusting, on-line method. Thus the optimization time is not critical.

In his interesting remarks Professor Feldbaum asks why our method is used, since man is a bad optimizer, and a self-adjusting model method is not an optimum parameter estimation method. Mr. Barker finds no advantage in our method over the use of simple feedback loops. If the problem were rapid on-line model optimization, we would agree with the two remarks, but when it comes to off-line model studies, which have widespread applications, we feel that our method has certain real advantages over automatic self-adjusting methods. When the total optimization time is not critical, man is in our opinion one of the best (and cheapest) optimizers one can have, due to his ability to adapt his strategy to the information gained during the optimization. As mentioned by Professor Sargent and Mr. Meyerhoff one may often encounter response surfaces with local minima or with

no convergence towards the desired optimum. In such cases all but the most elaborate automatic methods will fail, since they require monotone convergence from the starting point towards the optimum. But with our manual method, with graphical display of the response curves for  $\varepsilon$ , it is easy to get a graphical picture of a broad range of the response surface, by plotting a family of  $\varepsilon$  curves for certain values of all the model parameters. From a study of these curves is found the area within which the absolute optimum probably lies, and a cyclical parameter adjustment towards this optimum can then be started.

Other advantages of our method are: (a) no complex instrumentation of automatic optimizer; (b) the time-scaling provided with ISAC, and (c) easy and accurate graphical interpolation of minimum point of  $\varepsilon$  curves.

A requirement on the manual adjustment method is that the basic strategy be simple and straightforward, such as the cyclical parameter adjustment. Therefore, in reply to Mr. Kumar, the steepest descent method is not particularly well suited for manual optimization, and has not been tried by us.

The statement by Mr. Meyerhoff that manual adjustments require that the parameters be non-interacting is not correct, although the method of course works better if they are.

#### *Applications to Variable or Non-linear Systems*

In reply to Professor Sargent, we have not yet made any experiments with multi-input models but we believe that if the two model channels I and II converge independently towards reasonable results, the alternate adjustment scheme should converge towards an even better result.

The problem posed by Mr. Rakowski is an interesting but difficult one. With unknown structure there is probably no hope for the model

method, nor in the case of known structure with too many unknown parameters. For known structure, and with no more than four or five adjustable parameters, it might be possible to optimize these by including the complete system with all interconnections and feedback loops in the model. A practical limitation is that only three signals can be recorded simultaneously on ISAC.

With regard to non-linear systems, the model method is, in our opinion, one of the very few that lend themselves to determination of the form of non-linear functions and operators that appear in the system.

#### *Other Methods*

We want to express our gratitude for the valuable contributions received from Messrs. Brookes, Kumar, Mesch, Meyerhoff, Reinisch, Veltman and Young. They show that much work is now being done in this field, and that many alternative methods exist for solving the general problem of process parameter estimation.

The error criterion proposed by Mr. Mesch works well for one parameter, but fails in multi-parameter problems. The first-order error approximation then becomes

$$\overline{x(t) \cdot [z(t) - u(t)]} \approx \sum_{i=1}^n \Delta K_i \int_{-\infty}^{\infty} S_{xx}(\omega) \frac{\partial G}{\partial K_i} d\omega$$

showing that the adjustment of one  $\Delta K_i$  to zero does not give zero error measure unless all other  $\Delta K_i$  are zero, which is a trivial case. In general zero-seeking error criteria should be avoided, since they put certain strict requirements on the magnitudes of additional constant terms in the error expression. Extremum-seeking criteria are to be preferred, being independent of constant terms.

# Les Erreurs Systématiques et Aléatoires dans la Détermination Experimentale des Fonctions de Transfert

J. LOEB

## Summary

Mathematical methods are known by which records obtained in the course of normal plant operation yield the time constants of transfer functions. Following these methods one systematically tries the proposed solutions, with, as a criterion of accuracy, the minimization of a mean square error.

The paper shows that, if the noise which is added at the output of the element under study is correlated with the input signal, the time constants which minimize the error are shifted with respect to correct ones, hence a systematic error occurs. This precisely happens in feedback systems. Moreover, there exists a random error coming from noise fluctuations.

The paper also shows how to evaluate the systematic errors and gives some conditions to be fulfilled by the records so that time constants can be calculated with tolerable errors.

Cases dealt with are those of transfer functions:

$$\frac{e^{-\theta_p}}{1+Tp} \text{ and } \frac{1}{1+Tp+T^2p^2}$$

## Sommaire

On connaît les méthodes mathématiques au moyen desquelles les enregistrements obtenus en fonctionnement normal de l'usine fournissent les constantes de temps des fonctions de transfert.

Ces méthodes comportent l'essai systématique des solutions proposées, le criterium d'exactitude étant le minimum d'une erreur moyenne quadratique.

L'auteur montre que si le bruit qui vient s'ajouter à la sortie de l'élément étudié est corrélé avec le signal d'entrée, les constantes de temps qui donnent l'erreur minimale sont décalées par rapport aux vraies, d'où l'existence d'une erreur systématique. C'est ce qui se passe justement dans les systèmes à réaction.

De plus, il existe une erreur aléatoire due aux fluctuations du bruit.

On indique une façon d'évaluer l'erreur systématique, et les conditions pour que les enregistrements puissent servir au calcul des constantes de temps avec des erreurs admissibles.

Les cas traités sont ceux des fonctions de transfert:

$$\frac{e^{-\theta_p}}{1+Tp} \text{ et } \frac{1}{1+Tp+T^2p^2}$$

## Zusammenfassung

Es gibt mathematische Methoden zur Ermittlung der Zeitkonstanten der Übertragungsfunktion anhand von Aufzeichnungen der Meßwerte bei ungestörtem Betriebszustand. Aufgrund dieser Methoden probiert man systematisch eine Reihe von vorgeschlagenen Lösungen, wobei als Maß für die Genauigkeit der mittlere quadratische Fehler zum Minimum gemacht wird.

Dieser Beitrag zeigt: Wenn man das Rauschsignal, das am Ausgang des untersuchten Elementes hinzuaddiert wird, mit dem Eingangssignal korreliert, dann sind die Zeitkonstanten verschoben — was das Auftreten eines systematischen Fehlers erklärt. Genau das passiert in Regelungssystemen. Obendrein tritt ein Zufallsfehler auf, der durch Rauschschwankungen bedingt ist.

Dieser Aufsatz zeigt auch, wie der systematische Fehler ermittelt

werden kann, und nennt die Bedingungen, die die Aufzeichnungen erfüllen müssen, damit sie zur Berechnung der Zeitkonstanten mit zulässigen Fehlern dienen können.

Die hier behandelten Fälle betreffen folgende Arten von Übertragungsfunktionen:

$$\frac{e^{-\theta_p}}{1+Tp} \text{ und } \frac{1}{1+Tp+T^2p^2}$$

## Introduction

Le présent travail se situe dans le cadre des techniques indiquées par Milsum<sup>1</sup>.

On connaît la forme des fonctions de transfert, mais on ignore les valeurs des différentes constantes de temps et des différents coefficients.

Les grandeurs d'entrée et de sortie sont enregistrées au cours du fonctionnement normal de l'usine; on essaie systématiquement les valeurs numériques cherchées, en reconstruisant les grandeurs de sortie, et on adopte celles qui donnent une erreur moyenne quadratique minimale.

Une telle technique ne donnerait les résultats exacts (avec une erreur qui dépendrait seulement des appareils de mesure) que si le «bruit» était absent.

Par bruit nous désignons tout l'ensemble de perturbations qui agissent sur les grandeurs mesurées, en dehors de celle dont on cherche l'effet.

Cela pourra être l'action des autres variables de l'usine, ou celle d'éléments incontrôlables. Par exemple, si on cherche la façon dont la température de sortie d'un réacteur varie en fonction du débit d'un des constituants à l'entrée, l'effet des variations des autres constituants sera un bruit. Le refroidissement d'une colonne placée à l'air libre, sous l'effet du vent ou de la pluie, en sera un autre.

Des techniques que nous indiquons, en partant de l'exposé de Milsum, ne sont utiles que si réellement il y a un bruit important.

## L'Erreur systématique

Nous allons montrer que dans certains cas (que nous précisons) l'ensemble des valeurs des constantes de temps  $T_1, T_2 \dots$  qui minimise l'erreur moyenne quadratique ne se confond pas avec celui des vraies valeurs. Nous donnons ici le détail des calculs pour quatre cas simples.

### Cas de l'intégrateur simple

Soit un élément d'usine dont on sait qu'il est représentable par la fonction de transfert:

$$\frac{1}{1+Tp}$$

On ignore encore  $T$  qu'on va chercher à déduire des enregistrements suivants (Figure 1):

$f(t)$  grandeur d'entrée

$h(t)$  grandeur de sortie (brute)

$g'(t)$  grandeur de sortie, non chargée de bruit qui correspond à  $f(t)$  avec l'hypothèse  $T = T'$ ,  $T'$  étant la constante de temps essayée.

La sortie accessible  $h(t)$  est en fait la somme de:  $g(t)$  transformée de  $f(t)$  par  $1/(1+Tp)$  (inaccessible) et  $n(t)$  une perturbation aléatoire.

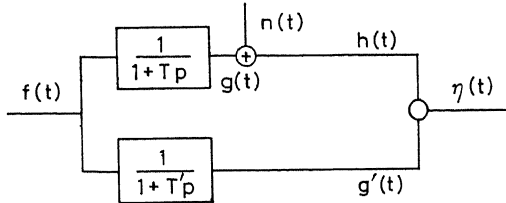


Figure 1

L'erreur (dont on cherche à réduire la valeur moyenne quadratique) est:

$$\eta = h - g'$$

Le schéma de la Figure 1 s'exprime algébriquement par:

$$\begin{aligned} (a) \quad f &= h - n + T(\dot{h} - \dot{n}) \\ (b) \quad f &= g' + T'\dot{g}' \end{aligned} \quad (1)$$

Le point sur une expression signifie «dérivation par rapport au temps».

Nous écrivons maintenant  $f$  au lieu de  $f(t)$  etc. ...

Les équations 1 (a) et (b) donnent:

$$\eta - n + T(\dot{h} - \dot{n}) - T'\dot{g}' = 0 \quad (2)$$

ou encore en ajoutant et retranchant  $T'\dot{h}$

$$\eta + T'\dot{\eta} = (T' - T)\dot{h} + n + T\dot{n} \quad (3)$$

Appliquons maintenant à (3) la transformation de Laplace, en notant  $L(f)$  la transformée.

$$(1 + T'p)L(\eta) = (T' - T)pL(h) + (1 + Tp)L(n) \quad (4)$$

Maintenant:

$$L(h) = \frac{L(f)}{1 + Tp} + L(n) \quad (5)$$

donc:

$$L(\eta) = \frac{(T' - T)pL(f)}{(1 + Tp)(1 + T'p)} + \frac{(T' - T)pL(n)}{1 + T'p} + \frac{(1 + Tp)L(n)}{1 + T'p}$$

ou:

$$L(\eta) = \frac{(T' - T)pL(f)}{(1 + Tp)(1 + T'p)} + L(n) \quad (6)$$

Appelons  $F(t)$  la transformée inverse du premier terme du second membre:

$$F(t) = \frac{pL(f)}{(1 + Tp)(1 + T'p)}$$

ou pratiquement  $F(t)$

$$F(t) = \frac{pL(f)}{(1 + Tp)^2} \quad (7)$$

Bien entendu,  $F(t) = 0$  si  $f$  est une constante.

Dans cette expression, figure l'inconnue  $T$  mais comme nous cherchons un ordre de grandeur, on pourra mettre au lieu de  $T$  et de  $T'$  une valeur du même ordre de grandeur que nous noterons  $T$  qu'on doit estimer *a priori*.

On a donc:

$$\eta = (T' - T)F(t) + n \quad (8)$$

Tirons de (8) par élévation au carré et calcul des moyennes:

$$\overline{\eta^2} = (T' - T)^2 \overline{F^2} + 2(T' - T)\overline{Fn} + \overline{n^2} \quad (9)$$

Ici  $\overline{Fn}$  est la valeur de la fonction d'intercorrélation  $\phi_{Fn}(\tau) = \overline{F(t - \tau)n(t)}$  entre  $F$  et  $n$  pour une valeur nulle du décalage  $\tau$ . Lorsque ce terme n'est pas nul, il y a une erreur systématique.

Nous savons déjà que  $\overline{\eta^2}$  va avoir un minimum positif.

En effet, le discriminant du second membre est:

$$(\overline{Fn})^2 - \overline{F^2} \cdot \overline{n^2}$$

Il est toujours négatif (inégalité de Schwarz).

Mais le minimum de  $\overline{\eta^2}$  quand on fait varier  $T'$  n'a plus lieu pour  $T' = T$ . Il a lieu pour:

$$T' = T - \frac{\overline{Fn}}{\overline{F^2}} \quad (10)$$

Or dans un système à réaction, le bruit  $n(t)$  se répercute sur la grandeur d'entrée  $f(t)$ .

Nous sommes ainsi placés dans le cas d'application de la formule (10).

Nous devons maintenant voir si cette erreur est importante et si un examen plus attentif des enregistrements ne va pas nous permettre de la déceler, puis de la corriger.

Remarquons d'abord que plus le système de retour entre  $n(t)$  et  $f(t)$  sera compliqué, plus de sources de bruits non corrélés avec  $n(t)$  viendront s'ajouter. Donc dans les installations importantes, il y a suffisamment de sources de bruit pour que la valeur moyenne  $\overline{Fn}$  soit petite.

On a d'ailleurs un moyen de s'en rendre compte par une analyse plus serrée des enregistrements.

Nous allons faire porter le travail de dépouillement sur deux intervalles de temps où la valeur moyenne quadratique  $\overline{F^2}$  n'est pas la même. Si en passant d'une portion d'enregistrement à l'autre,  $\overline{F^2}$  se trouve multipliée par un nombre  $\lambda^2$ , l'erreur sur  $T'$  tirée de (2) se trouve divisée par  $\lambda$ .

Ce résultat approximatif suppose que  $n(t)$  est peu corrélé avec  $F(t)$ , de telle sorte qu'on conservera le même  $n(T)$  dans les deux équations (1).

On ne trouvera donc pas la même valeur pour  $T'$  et les deux dépouillements donneront:

$$\begin{aligned} T'_1 &= T - \frac{\overline{Fn}^{T_1}}{\overline{F^2}} \\ T'_2 &= T - \frac{1}{\lambda} \frac{\overline{Fn}}{\overline{F^2}} \end{aligned} \quad (11)$$

$T'$  et  $T'_2$  ainsi que  $\lambda$  sont connues.

En éliminant  $\overline{F(t)n(t)}$  on trouve

$$T = \frac{\lambda T'_1 - T'_1}{\lambda - 1} \quad (12)$$

On trouvera toujours dans l'enregistrement deux périodes de temps pendant lesquelles les variations de  $f(t)$  ne sont pas les mêmes et le calcul (4) sera toujours possible. (Bien entendu si  $T'_1$  est voisin de  $T'_1$  il n'y a plus de problème:  $n(t)$  n'est pas corrélé avec  $F(t)$ ). De plus l'ensemble (11) donnera une valeur de  $\overline{F(t)n(t)}$  par élimination de  $T$ .

$$\overline{Fn} = \overline{F^2} \frac{\lambda}{1-\lambda} (T'_1 - T'_1) \quad (13)$$

Cas du retard pur (Figure 2)

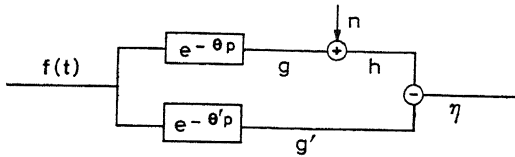


Figure 2

Si  $\theta$  est le retard inconnu,  $\theta'$  le retard essayé, les relations sont:

$$\eta = h - g' = f(t - \theta) - f(t - \theta') + n \quad (14)$$

ou encore

$$\eta = (\theta - \theta') f(t - \theta) + n \quad (15)$$

et

$$\overline{\eta^2} = (\theta - \theta')^2 \overline{f^2} + \overline{n^2} + 2(\theta - \theta') \overline{nf} \quad (16)$$

Il n'est plus nécessaire ici de préciser que les fonctions sont calculées pour l'instant  $t - \theta$ , car nous les supposons stationnaires. Comme précédemment la dérivée de  $\overline{\eta^2}$  par rapport à  $\theta - \theta'$  ne s'annule pas pour  $\theta = \theta'$  mais pour:

$$\theta' = \theta - \frac{\overline{fn}}{\overline{f^2}} \quad (17)$$

Cas où il y a un retard et une Intégration

La fonction de transfert essayée est:

$$\mathcal{E}(p) = \frac{e^{-\theta p}}{1 + T'p} \quad (18)$$

si bien que:

$$f(t) = g'(t - \theta') + T'g'(t - \theta') \quad (19)$$

avec toujours:

$$\eta = h - g'$$

Faisons dans (19) la substitution:

$$\begin{aligned} g'(t - \theta') &= g'(t - \theta) + (\theta - \theta') \dot{g}'(t - \theta) \\ \dot{g}'(t - \theta') &= \dot{g}'(t - \theta) + (\theta - \theta') \ddot{g}'(t - \theta) \end{aligned} \quad (20)$$

Retranchons (19) de (18) et remarquons que grâce à (20) toutes nos fonctions sont repérées à l'instant  $(t - \theta)$ . De plus ajoutons et retranchons  $T' \dot{h}$

$$0 = h - n - g' - (\theta - \theta') \dot{g}' + T'(\dot{h} - \dot{n})$$

$$- T'(\dot{g}' + (\theta - \theta') \ddot{g}') + T' \dot{h} - T' \dot{n}$$

d'où:

$$\eta + T \dot{\eta} = (T' - T) \dot{h} + n + T \dot{n} + (\theta - \theta') [\dot{g}' + T' \ddot{g}'] \quad (21)$$

ou encore à cause de (1)

$$\begin{aligned} \eta + \dot{\eta} [T' + (\theta - \theta')] + T'(\theta - \theta') \ddot{\eta} \\ = [T' - T + \theta - \theta'] \dot{h} + T'(\theta - \theta') \ddot{h} + n + T \dot{n} \end{aligned} \quad (22)$$

Appliquons à (22) la transformation de Laplace.

$$\begin{aligned} [1 + (T' + \theta - \theta')p + T'(\theta - \theta')p^2] L(\eta) \\ = [(T' - T + \theta - \theta')p + T'(\theta - \theta')p^2] L(h) + (1 + Tp) L(n) \end{aligned} \quad (23)$$

Puis comme:

$$L(h) = \frac{L(f)e^{-\theta p}}{1 + Tp} + L(n)$$

$$L(\eta) = \frac{L(f)e^{-\theta p} [(T' - T + \theta - \theta')p + T'(\theta - \theta')p^2]}{(1 + Tp)(1 + (T' + \theta - \theta')p + T'(\theta - \theta')p^2)} + L(n) \quad (24)$$

Les valeurs moyennes ne sont pas changées par la translation dans le temps  $\theta$  en raison des propriétés stationnaires des fonctions.

Pour évaluer les coefficients de  $T' - T$  et  $\theta - \theta'$  nous pourrions négliger  $\theta - \theta'$  au dénominateur de (24). Pour exprimer  $L(\eta)$  nous aurons besoin, en plus de  $F(t)$  définie en (7) de:

$$G(t) = \frac{L(f)p^2}{(1 + Tp)(1 + T'p)} \quad \text{ou} \quad G(t) = \frac{L(f)p^2}{(1 + Tp)^2} \quad (25)$$

Partons de (24) qui, avec les simplifications admises, et en négligeant le terme exponentiel qui ne fait que décaler la variable  $t$ , donne:

$$\eta = F(T' - T + \theta - \theta') + T'(\theta - \theta') + n \quad (26)$$

En élevant (26) au carré et en notant que:  $\overline{F\dot{F}} = 0$  il vient:

$$\begin{aligned} \overline{\eta^2} = \overline{F^2} [T' - T + \theta - \theta']^2 + \overline{F^2} T'^2 (\theta - \theta')^2 \\ + \overline{n^2} + 2\overline{Fn} [T' - T + \theta - \theta'] + 2\overline{Fn} T' (\theta - \theta') \end{aligned} \quad (27)$$

Posons:

$$x = T' - T$$

$$y = \theta' - \theta$$

En annulant les dérivées partielles

$$\frac{\partial \overline{\eta^2}}{\partial x} \quad \text{et} \quad \frac{\partial \overline{\eta^2}}{\partial y}$$

on a:

$$\begin{aligned} 2(x - y) \overline{F^2} + 2\overline{Fn} &= 0 \\ -2(x - y) \overline{F^2} - 2\overline{Fn} + 2\overline{F^2} T'^2 y - 2\overline{Fn} T' &= 0 \end{aligned} \quad (28)$$

On retrouve bien entendu  $x = y = 0$  si  $\overline{Fn} = 0$  et si  $\overline{F\dot{F}} = 0$ . S'il n'en est pas ainsi:

$$\begin{aligned} \theta' &= \theta + \frac{\overline{Fn}}{T' \overline{F^2}} \\ T' &= T + \frac{\overline{Fn}}{\overline{F^2} T'} - \frac{\overline{Fn}}{\overline{F^2}} \end{aligned} \quad (29)$$



Comme dans le cas que nous avons considéré précédemment, on lèvera l'incertitude en prenant deux enregistrements donnant des valeurs différentes pour  $\overline{F^2}$ .

#### Cas de la double intégration

La fonction de transfert est du type

$$\mathcal{E}(p) = \frac{1}{1 + Tp + \theta^2 p^2}$$

Les temps  $T$ ,  $\theta$  sont les inconnues du problème. La relation entre  $f$  et  $g$  est alors (en conservant les mêmes notations).

$$f = g + T\dot{g} + \theta^2 \ddot{g}'$$

ou

$$f = h - n + T(\dot{h} - \dot{n}) + \theta^2(\ddot{h} - \ddot{n}) \quad (30)$$

La fonction de transfert essayée donne:

$$f = g' + T'\dot{g}' + \theta'^2 \ddot{g}' \quad (31)$$

et

$$\eta - n + T(\dot{h} - \dot{n}) - T'\dot{g}' + \theta^2(\ddot{h} - \ddot{n}) - \theta'^2 \ddot{g}' = 0$$

Ajoutons et retranchons:

$$T'\dot{h} + \theta'^2 \ddot{h}$$

Il vient:

$$\eta + T'\dot{\eta} + \theta'^2 \ddot{\eta} = (T' - T)\dot{h} + (\theta'^2 - \theta^2)\ddot{h} + n + T\dot{n} + \theta^2 \ddot{n} \quad (32)$$

Appliquons la transformation de Laplace

$$(1 + T'p + \theta'^2 p^2)L(\eta) = (T' - T)pL(h) + (\theta'^2 - \theta^2)p^2L(h) + (1 + Tp + \theta^2 p^2)L(n)$$

Or:

$$L(h) = \frac{L(f)}{1 + Tp + \theta^2 p^2} + L(n)$$

Il vient:

$$L(\eta) = \frac{L(f) \cdot [(T' - T)p + (\theta'^2 - \theta^2)p^2]}{(1 + Tp + \theta^2 p^2)(1 + T'p + \theta'^2 p^2)} + \frac{(T' - T)p + (\theta'^2 - \theta^2)p^2 + 1 + Tp + \theta^2 p^2}{1 + T'p + \theta'^2 p^2} L(n)$$

qui se simplifie en:

$$L(\eta) = \frac{L(f)[(T' - T)p + (\theta'^2 - \theta^2)p^2]}{(1 + Tp + \theta^2 p^2)(1 + T'p + \theta'^2 p^2)} + L(n) \quad (33)$$

Posons cette fois-ci:

$$G(t) = \frac{L(f)p}{(1 + Tp + \theta^2 p^2)(1 + T'p + \theta'^2 p^2)}$$

ou pratiquement:

$$G(t) = \frac{L(f)p}{(1 + Tp + \theta^2 p^2)^2} \quad (34)$$

$$\eta = (T' - T)G + (\theta'^2 - \theta^2)\dot{G} + n \quad (35)$$

ou

$$\overline{\eta^2} = (T' - T)^2 \overline{G^2} + (\theta'^2 - \theta^2)^2 \overline{\dot{G}^2} + \overline{n^2} + 2(T' - T)\overline{Gn} + 2(\theta'^2 - \theta^2)\overline{\dot{G}n} \quad (36)$$

Toujours avec  $T' - T = x$  et  $\theta' - \theta = y$ , en égalant à zéro les dérivées partielles

$$\frac{\partial \overline{\eta^2}}{\partial x} \quad \text{et} \quad \frac{\partial \overline{\eta^2}}{\partial y}$$

et en supposant  $\theta' - \theta < \theta$  on obtient les valeurs en  $T'$  et  $\theta'$  qui rendent  $\overline{\eta^2}$  minimale.

$$T' = T - \frac{\overline{Gn}}{\overline{G^2}} \quad (37)$$

$$\theta' = \theta - \frac{\overline{Gn}}{2\theta \overline{G^2}}$$

Là encore, la technique reposant sur l'analyse de deux intervalles de temps s'appliquera.

#### Les erreurs aléatoires

Supposons maintenant que le bruit n'est pas corrélé avec l'entrée, c'est à dire qu'on a pu négliger ou corriger les termes tels que  $\overline{Fn}$ ; il restera pour  $\overline{\eta^2}$  des expressions de la forme  $A(T' - T)^2 + B$ . Le terme  $B$  contient  $\overline{n^2}$ . Si le bruit était une fonction aléatoire rigoureusement stationnaire, les essais fourniraient  $T'$  sans autre erreur.

En fait, il n'est pas prudent de considérer  $\overline{n^2}$  comme constant tout au long de l'enregistrement qui sert au dépouillement.

Une valeur essayée de  $T'$  conduit à une fonction  $g'(t)$  que l'on compare à la véritable fonction  $h(t)$ . Si le bruit n'est pas stationnaire, la valeur essayée  $T'$  (ou surtout  $\theta'$  s'il s'agit d'un retard pur) va nous conduire dans une région de la courbe  $h(t)$  qui ne sera pas la même que celle qui correspondra à la vraie valeur  $T$  ou  $\theta$ .

#### Evaluation des erreurs aléatoires sur $T$ et $T'$

Pour évaluer l'erreur qui serait provoquée par des fluctuations  $\delta \overline{n^2}$  de  $\overline{n^2}$ , nous allons exprimer  $\overline{\eta^2}$  en fonction de:

La fonction d'entrée  $f(t)$  ou de celles  $F$  et  $G$  qui s'en déduisent respectivement par les équations (7) et (25).

Les fluctuations du bruit. Ces expressions nous sont données par les calculs précédents.

*Cas d'une seule intégration* — Il suffit de faire dans (9)  $\overline{Fn} = 0$  ce qui donne

$$\overline{\eta^2} = (T' - T)^2 \overline{F^2} + \overline{n^2} \quad (38)$$

#### Cas d'un retard pur

$$\eta^2 = (\theta - \theta')^2 \overline{f^2} + \overline{n^2} \quad (39)$$

#### Retard et intégration

$$\overline{\eta^2} = \overline{F^2} (T' - T + \theta - \theta')^2 + \overline{F^2} T'^2 (\theta - \theta')^2 + \overline{n^2} \quad (40)$$

#### Double intégration

$$\overline{\eta^2} = (T' - T)^2 \overline{G^2} + (\theta'^2 - \theta^2)^2 \overline{\dot{G}^2} + \overline{n^2} \quad (41)$$

Au moment où on fera les tâtonnements sur  $T'$  et  $\theta'$  il est raisonnable de penser que  $\overline{n^2}$  aura bougé de  $\delta \overline{n^2}$ . On arrive ainsi aux 4 expressions de l'erreur aléatoire.

*Une intégration*

$$\Delta T \sim \left( \frac{\delta n^2}{F^2} \right)^{\frac{1}{2}} \quad (42)$$

*Retard pur*

$$\Delta \theta \sim \left( \frac{\delta n^2}{f^2} \right)^{\frac{1}{2}} \quad (43)$$

*Retard et intégration*

$$\Delta T + \Delta \theta \sim \left( \frac{\delta n^2}{F^2} \right)^{\frac{1}{2}} \quad (44)$$

$$\Delta \theta \sim \frac{1}{T^2} \left( \frac{\delta n^2}{F^2} \right)^{\frac{1}{2}}$$

*Double intégration*

$$\Delta T \sim \left( \frac{\delta n^2}{G^2} \right)^{\frac{1}{2}} \quad (45)$$

$$\Delta \theta \sim \frac{1}{4\theta^2} \left( \frac{\delta n^2}{G^2} \right)^{\frac{1}{2}}$$

*Conditions à remplir pour les enregistrements*

L'usage qu'on veut faire de la connaissance des fonctions de transfert impose une limite aux erreurs sur  $T$  et  $\theta$ . La connaissance des retards  $\theta$  déterminera la limite du spectre des perturbations entre lesquelles l'usine peut encore être « contrôlable ». Celle des constantes  $T$  déterminera la structure (et les valeurs numériques des composants) des réseaux correcteurs à placer dans la boucle de réaction.

Or il se peut que les amplitudes des variations des grandeurs telles que  $f(t)$  ne soient pas assez grandes pour que les équations (42) à (45) conduisent à des valeurs tolérables des erreurs  $\Delta T$  et  $\Delta \theta$ .

Les fonctions  $F$  et  $G$  sont, certes, construites au moyen des valeurs recherchées  $T$  et  $\theta$  mais on en connaît l'ordre de grandeur, et les équations (42) à (45) peuvent servir de guides pour décider si on entreprendra la recherche des valeurs minimales de  $\eta^2$  ou pour choisir dans les enregistrements les parties les plus favorables.

**Bibliographie**

- <sup>1</sup> MILSUM, J. H. Transfer function discovery on an analog computer. *I. S. A. Commun.* No. 55-59, 1959. Chicago

**DISCUSSION**

P. EYKHOFF, *Technological University, Department of Electrical Engineering, Kanaalweg 2B, Delft, Netherlands*

Dr. Loeb's paper is devoted to the interesting problem of process-parameter estimation under the condition that the process-input signal and the additive noise are correlated.

It may be of interest to point out the relation between the operations used in the paper to derive an expression like eqn (8) of the paper and the idea of parameter-influence coefficients, introduced by Meissinger<sup>1, 2</sup>.

These parameter-influence coefficients are the partial derivatives of problem variables with respect to system parameters.

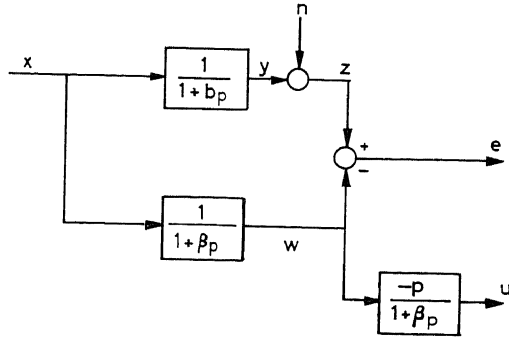


Figure A

Using the notation indicated in Figure A

$$b\dot{y} + y = x \quad (1)$$

$$\beta\dot{w} + w = x \quad (2)$$

$$e = y - w + n \quad (3)$$

Following Meissinger, the model output  $w$  can be considered as a function of two variables;  $w = w(t, \beta)$ , therefore

$$w(t, \beta + \Delta\beta) = w(t, \beta) + \left( \frac{\partial w}{\partial \beta} \right)_\beta \Delta\beta + \left( \frac{\partial^2 w}{\partial \beta^2} \right)_\beta \left( \frac{\Delta\beta}{2!} \right)^2 + \dots \quad (4)$$

If  $b = \beta + \Delta\beta$  then  $y = w(t, \beta + \Delta\beta)$  and

$$y - w = \left( \frac{\partial w}{\partial \beta} \right)_\beta (b - \beta) + \left( \frac{\partial^2 w}{\partial \beta^2} \right)_\beta \left( \frac{b - \beta}{2!} \right)^2 + \dots \quad (5)$$

For a small difference  $b - \beta$  only the first term at the right-hand side of eqn (5) is used. Eqn (5) substituted in (3) then leads to

$$e \approx \left( \frac{\partial w}{\partial \beta} \right)_\beta (b - \beta) + n \quad (6)$$

which is the same as eqn (8) of the paper.

The parameter influence coefficient can be determined by partial differentiation of eqn (2) with respect to  $\beta$ :

$$\dot{u} + u = -\dot{w} \quad (7)$$

where

$$u = \frac{\partial w}{\partial \beta} \quad (8)$$

This time-function can be generated as is indicated in Figure A. It will be clear that it is essentially the same as the function  $F(t)$  introduced by eqn (7) of the paper.

This approach has the advantage that the higher-order parameter influence coefficients in eqn (5) indicate over what parameter-interval the approximation of eqn (6) holds.

Minimization of  $e^2$  with respect to  $\beta$  leads to:

$$0 = \frac{\partial e^2}{\partial \beta} = 2e \frac{\partial e}{\partial \beta} = -2e \frac{\partial w}{\partial \beta} = -2eu \quad (9)$$

With eqns (6) and (8) this leads to the same condition as given by eqn (10) of the paper.

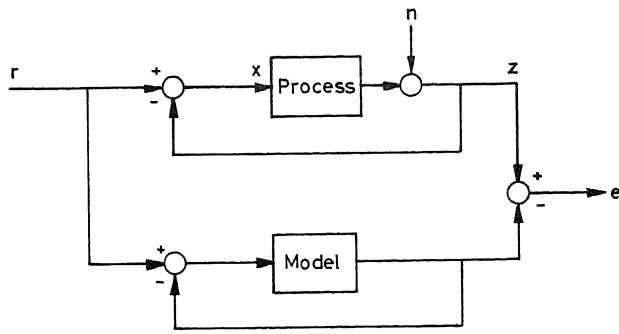


Figure B

The paper gives an interesting method of correction for the adverse effect of the correlation between process input signal and additive noise. One interesting point that this raises in my mind is whether a scheme analogous to *Figure B* could always be used and for what reason.

#### References

- <sup>1</sup> MEISSINGER, H. F. The use of parameter influence coefficients in computer analysis of dynamic systems. *Proc. West. Joint Computer Conf.*, San Francisco (May 1960), 181-192
- <sup>2</sup> MEISSINGER, H. F. Parameter influence coefficients and weighting functions applied to perturbation analysis of dynamic systems. *Proc. Third Internat. Congr. Analog Computation*, Opatija, Yugoslavia (Sept. 1961)

# The Applicability of the Relay Correlator and the Polarity Coincidence Correlator in Automatic Control

B. P. Th. VELTMAN and A. VAN DEN BOS

## Summary

The increasing need for a statistical evaluation of signals for adaptive and optimizing control purposes makes simple, reliable correlating devices desirable. In many cases the relay correlator or the polarity coincidence correlator can provide satisfactory answers.

These methods are not restricted to Gaussian signals, but can be applied to any type of joint distribution function of the signals by the addition of a special auxiliary signal. It is shown that the increase in statistical inaccuracy with this procedure can be diminished by taking more samples within the given observation time of the signals. A generator for the auxiliary signal is described.

The applicability of the polarity coincidence method on Gaussian signals may be restricted by the presence of disturbing signals. It is proved that in certain cases the zero crossings and the extreme values of the correlation function do not shift with respect to the  $\tau$  axis. Examples of the cross-correlation of Gaussian signals disturbed by noise and of the detection of sinusoidal signals out of noise are given. Finally the capabilities of the polarity coincidence method are illustrated with the measurement of the dynamic behaviour of a system in the frequency domain. Two correlating devices whose usefulness has been proved are described in an Appendix.

## Sommaire

Des corrélateurs simples et fiables sont de plus en plus souhaités en raison de la nécessité d'une évaluation statistique des signaux dans les commandes adaptatives et optimalisantes. Dans bien des cas, les corrélateurs à relais ou corrélateurs à coïncidence de polarité donnent des résultats satisfaisants.

Ces corrélateurs peuvent s'appliquer non seulement à des signaux gaussiens, mais à n'importe quel type de fonction de distribution jointe des signaux par l'addition d'un signal auxiliaire spécial. On montre que l'imprécision statistique peut être réduite en prenant plus d'échantillons à l'intérieur de la durée d'observation des signaux. On décrit un générateur de signal auxiliaire.

L'efficacité de ces corrélateurs dans le cas des signaux gaussiens peut être diminuée par la présence de signaux perturbateurs. On montre que dans certains cas la valeur pour  $\tau = 0$  et les valeurs extrêmes de la fonction d'autocorrélation ne changent pas de position le long de l'axe des  $\tau$ . Quelques exemples sont donnés. Finalement on montre que la méthode à coïncidence de polarité permet de mesurer le comportement fréquentiel d'un système. Deux corrélateurs éprouvés sont décrits.

## Zusammenfassung

Der wachsende Bedarf nach einer statistischen Auswertung von Signalen für selbststellende und optimale Regelungen erfordert einfache und zuverlässige Korrelatoren. In vielen Fällen liefert der Relais- oder Polaritäts-Koinzidenz-Korrelator befriedigende Lösungen.

Diese Methoden sind nicht auf Signale mit Normalverteilung beschränkt, sondern sie finden auch auf Signale mit einer anderen Verteilungsfunktion Anwendung, indem man ein spezielles Hilfssignal hinzuaddiert. Es zeigt sich, daß bei diesem Verfahren gesteigerte statistische Streuung dadurch herabgesetzt wird, daß man eine größere Anzahl von Proben des Signals in dem gegebenen Beobachtungszeitraum entnimmt. Ein Generator für das Hilfssignal wird beschrieben.

Die Anwendbarkeit der Polaritätskorrelation auf normal verteilte Signale kann durch vorhandene Störsignale beschränkt werden. Es wird bewiesen, daß sich in bestimmten Fällen die Nulldurchgänge und die Extremwerte der Korrelationsfunktion bezüglich der  $\tau$ -Achse nicht verschieben. Beispiele zeigen die Kreuzkorrelation von normal verteilten Signalen, die durch Rauschsignale gestört sind, und die Auffindung von Sinussignalen aus Störgeräuschen.

Schließlich werden die Möglichkeiten der Polaritätskorrelation erläutert, die sich anhand der Messung des dynamischen Verhaltens eines Systems im Frequenzbereich ergeben.

Der Anhang enthält die Beschreibung zweier Korrelatoren, die sich in der Praxis bewährten.

## 1. Introduction

A true estimate, for a finite observation time  $T$ , of the normalized correlation function  $C(\tau)$  between two signals  $x_1(t)$  and  $x_2(t)$  can be obtained by the operation

$$c(\tau) = \frac{\langle x_1(t) \cdot x_2(t-\tau) \rangle_T}{\langle x_1(t) \cdot x_2(t) \rangle_T} \quad (1)$$

An estimate of the normalized relay correlation function  $R(\tau)$  is

$$r(\tau) = \frac{\langle x_1(t) \cdot \text{sgn } x_2(t-\tau) \rangle_T}{\langle |x_1(t)| \rangle_T} \quad (2)$$

An estimate of the polarity coincidence correlation function\*  $P(\tau)$  is

$$p(\tau) = \langle \text{sgn } x_1(t) \cdot \text{sgn } x_2(t-\tau) \rangle_T \quad (3)$$

'Sgn' means that the polarity compared to the average value of the signal has to be taken,  $\tau$  is a timeshift and  $\langle \rangle_T$  is the average value over the observation time  $T$ .

It is obvious that the instrumentation for simulating eqns (2) and (3) is much simpler than for eqn (1). Timeshift, multiplication and integration are easily instrumentated with 1 bit signals, as is pointed out in the Appendix. Several less rigorous simplifications in operation are possible with acceptable extensions as to the computer complexity<sup>1</sup>; however, the important advantages with the measuring and transportation of 1 bit signals are lost with these other methods. Although several authors in the past decade report these significant simplifications, the n.r.c.f. and the p.c.c.f. methods are not in common use in control engineering.

Intuitively one argues that the simplifications must be paid for by a loss of information. The relative short observation times that are usual in control engineering do not permit any waste

\* Normalized correlation function will now be abbreviated by n.c.f.; normalized relay correlation function by n.r.c.f.; polarity coincidence correlation function by p.c.c.f.

of information. It can be shown, however, that this intuitive reasoning does not hold in a number of cases.

The following relations have been proved<sup>1-4</sup>.

(i) Calculation of the normalized correlation function from the normalized relay correlation function or from the polarity correlation function for stationary Gaussian ergodic signals with zero means:

$$\text{n.r.c.f.} \quad C(\tau) = R(\tau) \quad (4)$$

$$\text{p.c.c.f.} \quad C(\tau) = \sin \left[ \frac{\pi}{2} \cdot P(\tau) \right] \quad (5)$$

$$\lim_{T \rightarrow \infty} r_T(\tau) = R(\tau) \quad (6)$$

$$\lim_{T \rightarrow \infty} p_T(\tau) = P(\tau) \quad (7)$$

(ii) The variances in the outcome of eqns (2) and (3) are of the same order of magnitude as those of eqn (1) for equal observation time  $T$ . In all cases the variance depends upon the fourth product moment of the signals (see also McFadden<sup>5</sup> and Fuller<sup>6</sup>).

(iii) The n.r.c.f. and p.c.c.f. methods are applicable to any ergodic process if a special auxiliary signal is added before the polarity is taken.

$$\text{n.r.c.f.} \quad C(\tau) = \frac{A_0}{\sigma_0} \cdot R(\tau) \quad (8)$$

$$\text{p.c.c.f.} \quad C(\tau) = \frac{A_0 A_1}{\sigma_0 \sigma_1} \cdot P(\tau) \quad (9)$$

$\sigma^2$  are the variances of the signals that have to be correlated. The auxiliary signal must have a uniform amplitude density function between the extreme values  $A$  of the signal. The correlation function of the auxiliary signal must be an approximation of the  $\delta$  function.

(iv) With eqns (8) and (9) the variances in the estimation of  $C(\tau)$  can be significantly larger than an estimation according to eqn (1).

From statement (ii) it can be concluded (and this has been extensively verified by experiments) that for Gaussian signals the n.r.c.f. and p.c.c.f. methods are not merely 'poor man's solutions', but they have almost the same power as the n.c.f. method.

According to (i) and (ii) it seems to be of no use with Gaussian signals, to apply more complicated devices than the polarity coincidence correlator. Restrictions must be made, however, when disturbing signals are present. Suppose that the correlation between signals  $x_0(t) + x_n(t)$  and  $x_i(t)$  is measured (Figure 1); the correlation between  $x_i(t)$  and  $x_0(t)$  should be established.

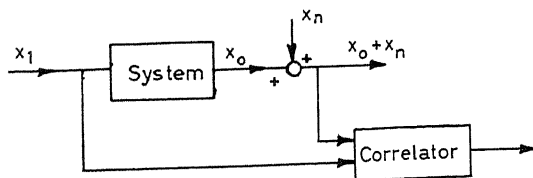


Figure 1. Measurement of the cross-correlation function between  $x_i(t)$  and  $x_0(t)$  if a disturbing signal  $x_n(t)$  is present

The presence of signal  $x_n(t)$  has no implements for the n.c.f. and the n.r.c.f. method.

According to eqn (4)  $\langle \text{sqn } x_i(t - \tau) \cdot [x_0(t) + x_n(t)] \rangle_T$  is proportional to  $\langle x_i(t - \tau) \cdot [x_0(t) + x_n(t)] \rangle_T$  which may be replaced by

$$\langle x_i(t - \tau) \cdot x_0(t) \rangle_T + \langle x_i(t - \tau) \cdot x_n(t) \rangle_T$$

This additive rule does not hold for the polarity coincidence method, due to the sine relation (5). It is shown in Section 3 of this paper under what conditions the p.c.c.f. method is still useful with added disturbances. According to (iv) above, the application of the n.r.c.f. and the p.c.c.f. seems to meet restrictions for non-Gaussian signals, if only a short observation time is permitted. In Section 2 it is shown that restriction (iv) can be weakened by increasing the number of samples within the given observation time. A generator for the auxiliary signal, that will permit such an increase, is described.

In Section 4 the capabilities of the p.c.c.f. are illustrated with an unusual method for the on-line measurement of some dynamic parameters of a system in the frequency domain.

## 2 (a). The Variance when the Auxiliary Signal is Added

The addition of the auxiliary signals makes the n.r.c.f. and the p.c.c.f. method applicable to any type of joint distribution function. It should be noted that it is not necessary to add different auxiliary signals; if the same auxiliary signal is added to all signals (e.g. with multiple correlations) it only makes the values of the correlation functions for  $\tau = 0$  undetermined. According to Veltman and Kwakernaak<sup>2</sup>, the variance in the p.c.c.f. when auxiliary signals are added is, for a sufficiently large time shift, approximated by

$$\text{var } p(\tau) = \frac{\sigma_0^2 \sigma_1^2}{A_0^2 A_1^2} \cdot \text{var } c(\tau) + \frac{1}{N} \left( 1 - \frac{\sigma_0^2 \sigma_1^2}{A_0^2 A_1^2} \right) \quad (10)$$

$\sigma^2$  are the variances of the signals;  $A$ 's are the extreme values of the signals.

$\text{var } c(\tau)$  is inversely proportional to the observation time  $T$ .  $N$  is the number of samples within the time interval  $T$ . It would be preferable to use in (10) the bandwidth of the auxiliary signal instead of the number of independent samples  $N$  within time  $T$ . A description of the constraints to the power spectrum of the auxiliary signal equivalent to 'succeeding samples must be independent' is, however, rather involved.

The right-hand side of eqn (10) can be divided into a contribution from the statistical properties of the signals themselves, and one introduced by the auxiliary signals. The comparison of eqn (9) with eqn (10) shows that, although the measured value  $p(\tau)$  is smaller than the wanted value  $c(\tau)$ , the relative inaccuracy due to the statistical properties of the signal alone is unaffected.

An increase in the number of samples  $N$  will decrease the statistical errors due to the addition of the auxiliary signals (provided that succeeding samples of the auxiliary signals remain independent). The question is, for what value of  $N$  is the influence of the auxiliary signal acceptable? 'Acceptable' can be expressed in terms of relative contribution to the total variance of the first and second terms in the right-hand side of eqn (10).

The uncertainty of an estimate can be described by

$$[\text{var } c(\tau)]^{\frac{1}{2}} / c(\tau).$$

Suppose

$$\frac{[\text{var } p(\tau)]^{\frac{1}{2}}}{p(\tau)} = \beta \cdot \frac{[\text{var } c(\tau)]^{\frac{1}{2}}}{c(\tau)}$$

with  $\beta$  as a measure for the increase in uncertainty with the addition of the auxiliary signal.

The substitution of eqns (9) and (10) in this expression, and putting  $\sigma_0/A_0 = \sigma/A_1 = \alpha$  results in:

$$\beta = \left(1 + \frac{(1 - \alpha^4)}{\alpha^4 N \text{var } c(\tau)}\right)^{\frac{1}{2}}$$

or

$$N = \frac{1 - \alpha}{\alpha^4 (\beta^2 - 1) \text{var } c(\tau)} \quad (11)$$

Realistic values of the parameters are  $\alpha = 1/3$ ;  $\text{var } c(\tau) = 0.001$  for  $N = 10^4$  (this strongly depends on the form of the correlation function). This means that the addition of the auxiliary signal leads to  $\beta = 3$ . To reach the value  $\beta = 1.5$ , more than six times as many samples are needed. If  $\text{var } C(\tau) = 10^{-3}$  for  $N = 10^5$ , the increase in inaccuracy with the addition of the auxiliary signal is only 35 per cent.

## 2 (b). A Generator for the Auxiliary Signal

To keep succeeding samples of the auxiliary signals independent, a high frequency generator for a random signal with uniform probability density function is needed. It is proved by Wonham and Fuller<sup>7</sup> that signals with a variety in probability density functions can be generated from a binary signal, in which the number of transients in a certain time interval obeys a Poisson distribution, by passing this binary signal through a first-order low pass filter.

It is usual to derive the Poisson binary signal from a radioactive source. The occurrence of a count from a Geiger-Müller tube or a scintillation crystal triggers a bistable circuit. Another approach, which leads to simpler equipment, is the derivation of the binary signal from Gaussian noise. The problem is how to shape the frequency spectrum of a gaussian signal in such a way that the distribution of the number of zero crossings of the Gaussian signal in a certain time interval can be approached by the Poisson formula. It may be assumed<sup>5</sup> that this is approximately achieved when the second moment of the probability density function of the clipped Gaussian noise equals the second moment of the Poisson on-off signal. This implies that the p.c.c.f. of the shaped Gaussian signal should be identical to the n.c.f. of the Poisson signal.

The n.c.f. of the Poisson signal is<sup>9</sup>

$$C(\tau) = e^{-2\lambda|\tau|} \quad (12)$$

in which  $\lambda$  is the Poisson parameter. The n.c.f. of the shaped Gaussian signal must be, according to eqn (5),

$$C(\tau) = \sin \left[ \frac{\pi}{2} \cdot e^{-2\lambda|\tau|} \right] \quad (13)$$

If (13) is developed into a McLaurin series and Fourier transformed to the frequency domain, it follows that the power spectrum of the shaped Gaussian signal is

$$|X(\omega)|^2 = \sigma^2 \cdot \sum_{n=0}^{\infty} \frac{2 \cdot (-1)^n \cdot \left(\frac{\pi}{2}\right)^{2n+1}}{(2n)!} \cdot \frac{2\lambda}{(2n+1)^2 (2\lambda)^2 + \omega^2} \quad (14)$$

which is a non-negative function.

Eqn (14) can be approximated by the first three terms of the series. After some modifications in the expression it follows

$$|X(\omega)|^2 = \sigma^2 \cdot \frac{2\pi\lambda(3145\lambda^4 + 18.2\lambda^2\omega^2 + 0.024\omega^4)}{(4\lambda^2 + \omega^2)(36\lambda^2 + \omega^2)(100\lambda^2 + \omega^2)} \quad (15)$$

which is realizable with operational amplifiers and RC circuits acting on wide band noise. The procedure gives a signal which, after clipping, has an auto-correlation function of the form  $e^{-2\lambda|\tau|}$

Experimentally it is established that simpler filters can be applied to the Gaussian signal, which also result in a signal with uniform amplitude density function after clipping and first-order low pass filtering. Satisfactory results are obtained by shaping wide-band noise with a first-order lag filter. The time constant of the second first-order filter must be about three times as large as the time constant of the first one to obtain a linear cumulative probability function. Figure 2 shows some results. It should be noted that succeeding intervals between zero crossings of

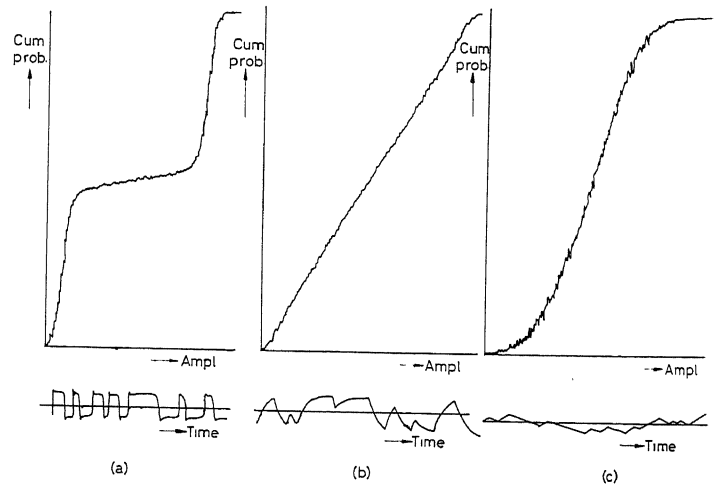


Figure 2. Cumulative probability functions of signals that are obtained by first-order lag filtering (time constant  $\theta_2$ ) of a clipped Gaussian signal. The Gaussian signal is obtained by first-order lag filtering (time constant  $\theta_1$ ) of wide band Gaussian noise. Samples of the corresponding signals for  $\theta_2/\theta_1 = 0$  (a),  $\theta_2/\theta_1 = 3$  (b) and  $\theta_2/\theta_1 = 30$  (c) are given

Gaussian signals are not independent. No true Poisson on-off signal can be obtained by filtering and clipping a Gaussian signal; although first and second moments of the signals are made the same, the higher moments must differ<sup>5</sup>.

## 3. The Polarity Coincidence Correlation Method Applied to Disturbed Signals

### (a) Cross-correlation between Gaussian Signals under the Presence of Added Noise

To be able to interpret the result of the polarity coincidence method in this case it would be necessary to know the joint probability density function between the input  $x_i(t)$  and the output  $x_0(t) + x_n(t)$  in Figure 1. This is not possible in general. A sophisticated method can be applied, however, for determining the average probability that both input and disturbed output have the same polarity<sup>8</sup>.

The joint probability density function between the undisturbed signals is integrated over an interval with the disturbing signal as a time-varying boundary. Averaging out in the time after integration results in an expression for the average probability for equal polarity of the disturbed signals. For Gaussian signals  $x_i(t)$  and  $x_0(t)$ , with zero mean and variances  $\sigma^2$ , the procedure is as follows. The joint probability of the undisturbed signals is

$$\frac{1}{2\pi\sigma_i\sigma_0(1-C^2)^{\frac{1}{2}}} \cdot \exp\left[\frac{x_i^2 - 2Cx_ix_0 + x_0^2}{2(1-C^2)\sigma_i\sigma_0}\right] \cdot dx_0 \cdot dx_i \quad (16)$$

in which  $C$  is the normalized correlation coefficient. If (16) is integrated over the interval  $[-n(t), \infty; 0, \infty]$  the probability for coinciding positive polarities is calculated from

$$P_{++} = \frac{1}{2\pi\sigma_i\sigma_0(1-C^2)^{\frac{1}{2}}} \int_{-n(t)}^{\infty} \int_0^{\infty} \frac{x_i^2 - 2Cx_ix_0 + x_0^2}{2\sigma_i\sigma_0(1-C^2)} \cdot dx_0 \cdot dx_i \quad (17)$$

By series development of this integral with respect to  $n(t)$  [assuming that the odd moments of the probability density function of  $n(t)$  are zero] and averaging out in time, it is found that the p.c.c.f. (which is equal to  $\langle 4P_{++} - 1 \rangle$ ) is given by

$$P(\tau) = \frac{2}{\pi} \left[ \arcsin C(\tau) - \frac{C(\tau)}{2!(1-C^2(\tau))^{\frac{1}{2}}} \cdot \frac{\langle n^2(t) \rangle}{\sigma_1^2} + \frac{C(\tau)(3-2C(\tau))^2}{4!(1-C^2(\tau))^{\frac{3}{2}}} \cdot \frac{\langle n^4(t) \rangle}{\sigma_1^4} + \dots \right] \quad (18)$$

This is identical to (5) if  $n(t) = 0$ ; (5) forms a good approximation for (18) if  $\langle n^2(t) \rangle / \sigma^2 < 0.1$ . It is evident that  $P(\tau)$  will be zero if  $C(\tau)$  is zero. Thus the zero crossings of the n.c.f. and the p.c.c.f. are identical. Differentiation of (18) with respect to  $\tau$  and substituting  $dC/d\tau = 0$  shows that the abscissae of the extreme values of the n.c.f. and the p.c.c.f. are the same. Figure 3 illustrates these statements with some measurements.

The applicability of the polarity coincidence method seems rather limited with disturbed signals. Interesting information about a system's dynamic behaviour can, however, be obtained from the shifts in zero crossings and extrema between the input autocorrelation and the input-output cross-correlation function, as is shown in Section 4.

#### (b) Restoring Sinusoidal Signals out of Added Gaussian Noise

This problem can be tackled in the same way as the procedure described in Section 3 (a). In this case  $P_{++}$  is given by a time varying integral of the bivariate probability density function of the gaussian noise (zero mean and variance  $\sigma_n^2$ ):

$$\text{Eqn (19)}^*$$

Series development with respect to  $a \sin \omega t$  and averaging out in time results in

$$^* \text{Eqn (19): } P_{++} = \frac{1}{2\pi\sigma_n^2(1-C^2)^{\frac{1}{2}}} \int_{-a \sin \omega t}^{\infty} \int_{-a \sin \omega(t-\tau)}^{\infty} e^{-\frac{x_n^2 - 2Cx_nx_{n\tau} + x_{n\tau}^2}{2\sigma_n^2(1-C^2)}} \cdot dx_n dx_{n\tau} \quad (19)$$

$$P(\tau) = \frac{2}{\pi} \left[ \arcsin C(\tau) + \frac{\cos \omega\tau - C(\tau)}{(1-C^2(\tau))^{\frac{1}{2}}} \cdot S + \frac{\cos \omega\tau - C(\tau)}{(1-C^2(\tau))^{\frac{3}{2}}} [C(\tau) \cos \omega\tau + C^2(\tau) - 2] \frac{S^2}{4} + \dots \right] \quad (20)$$

with  $S = a^2/2\sigma^2$ .

If  $C(\tau) = 0$  (for sufficiently large  $\tau$ ) eqn (20) reduces to

$$P(\tau) = \frac{2}{\pi} \cos \omega\tau \left[ S - \frac{S^2}{2} + \frac{S^3}{36} (\cos^2 \omega\tau + 6) - \frac{S^4}{48} (\cos^2 \omega\tau + 2) + \dots \right] \quad (21)$$

The n.c.f. for large values of  $\tau$  is

$$C(\tau) = \frac{S}{1+S} \cdot \cos \omega\tau \quad (22)$$

Comparison of eqn (21) with eqn (22) shows that the zero crossings of the n.c.f. and the p.c.c.f. are identical. Taking the first derivative in eqn (21), and substituting  $\omega\tau = (2k+1)\pi/2$  gives

$$P^1(\tau) = (-1)^{k+1} \cdot \frac{2\omega}{\pi} \left( S - \frac{S^2}{2} + \frac{S^3}{6} - \frac{S^4}{24} + \frac{S^5}{120} - \dots \right)$$

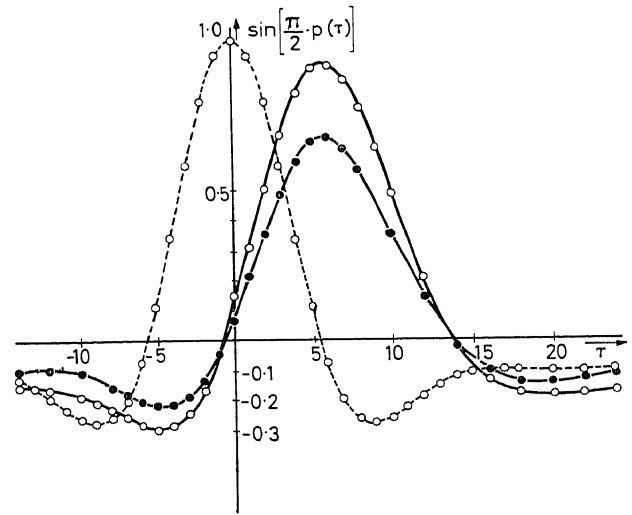


Figure 3. Measured cross-correlation functions  $\sin[\pi/2 \cdot p(\tau)]$  between input and output of a third-order system. The position of the zero crossings and the extrema is unaffected when noise is added. The third-order system consists of three equal time constants of 0.1 sec. The input signal is a Gaussian random signal with a flat spectrum between 0.2 and 2 c/sec (dotted line is the input correlation function)

○—○—○ without noise

●—●—● signal-to-noise power ratio 1.55; the noise has a flat spectrum between 0.2 and 2 c/sec

Sampling distance is 44 msec; observation time  $10^5$  samples = 4,400 sec

which is equal to

$$\frac{dP(\tau)}{d\tau} = (-1)^{k+1} \cdot \frac{2\omega}{\pi} [1 - e^{-S}] \quad (23)$$

The abscissae of the extreme values are thus identical with the n.c.f. and the p.c.f. if  $\tau$  is sufficiently large. It should be noted that the signal to noise power ratio  $S$  can be estimated from the intersection point of the tangents in succeeding zero crossings of the p.c.f., according to eqn (23). Figure 4 shows the experimental verification of these derivations.

#### 4. Measurements of the Dynamic Properties of a System with the Polarity Coincidence Method

Measurements for evaluating dynamic parameters of systems are presented as an illustration of the applicability of the p.c.f. in automatic control. It should be noted that any estimate of a correlation function, however bad it may be, always shows a very true appearance. This deceitful result is due to the fact that a correlation function is not a set of independent correlation coefficients. There exists a strong dependency between the items of a time series representing a continuous signal. This dependency causes a correlation between the variances in neighbouring points of the correlation function; apparently periodic components, that are not present in the signal, may be introduced by this phenomenon.

Thus it is necessary to calculate confidence limits with each measured correlation function to give an adequate interpretation. The measurements presented in this article are always extended

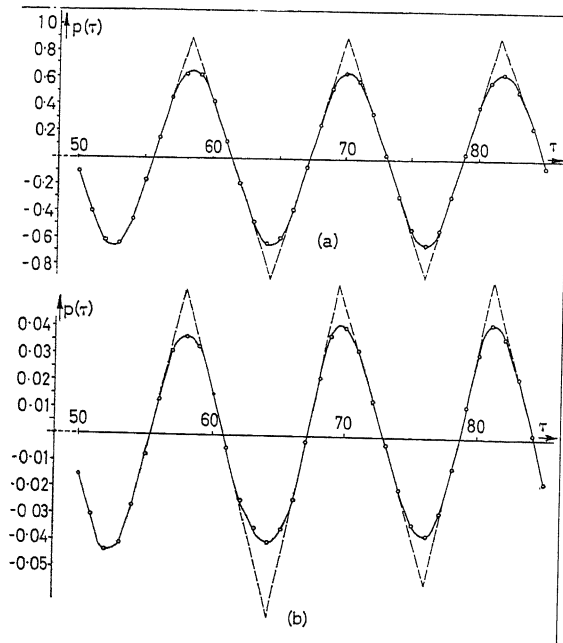


Figure 4. Measured polarity correlation functions of a sinusoidal signal disturbed by Gaussian noise, for large values of the timeshift  $\tau$ . It can be shown from these figures that zero crossings and extrema are not affected by the added noise. Sinusoidal signal, approx. 2 c/sec; sampling distance, 44 msec; observation time,  $10^5$  samples (about 8000 periods of the signal); noise bandwidth, 0.02–5 c/sec; signal-to-noise ratios measured: (a) 2.90 (b) 0.08; calculated from these graphs according to eqn (23), (a) 2.30 (b) 0.06

over a very long period; in most cases the confidence limits (95 per cent certainty that the true value is within the given interval) approach line thickness in the figures. Figure 5 gives the measured cross-correlation function of the input and output

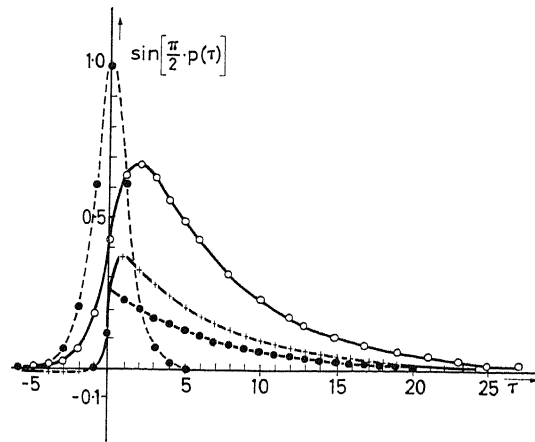


Figure 5. Measured cross-correlation function  $\sin(\pi/2) \cdot p(\tau)$  between input and output signal of a first-order low pass filter for different bandwidths of the input signal. Time constant of the filter 0.3 sec; sampling distance 44 msec; observation time  $10^5$  samples.

Bandwidth of the input signal

○—○—○ 0.5 c/sec (auto-correlation function of this signal is

+—+—+ 0.25 c/sec presented as a dotted line)

●—●—● 0.100 c/sec

No extra noise is added

signal of a first-order low pass filter, the input signal being a Gaussian random signal with a flat power spectrum. No extra noise is added. Deconvolution of the cross-correlation function with the auto-correlation function produces the pulse response of the system. The tedious deconvolution can be simplified by using wide-band noise for testing the system. The measurements of Figure 5 show indeed that the cross-correlation function approaches the system's pulse response if the bandwidth of the input signal is increased. It should be noted, however, that the amplitude of the measured normalized cross-correlation function deteriorates with an increasing bandwidth of the input signal. Only a small part of the input power spectrum appears at the output, making the value of the normalized cross-correlation small. This inaccuracy is not due to statistical inference but to the fact that two large numbers, the integrals of the positive and the negative result of the multiplication, are subtracted, giving a large measuring error in the difference.

More accurate results are obtained if the signals that have to be correlated cover the same part in the frequency domain. Suppose, for example, that input and output of the system are filtered by two identical small bandwidth filters (Figure 6). The output of the filters is shown in Figure 7. The wave-form of the filtered output signal seems to be shifted in time with respect to the filtered input signal. This is due to the almost linear phase shift and constant gain of the system for a small bandwidth signal.

The correlation functions describing the small bandwidth behaviour of a third-order system, measured by the polarity method, are given in Figure 8. Indeed the phase-shift of the



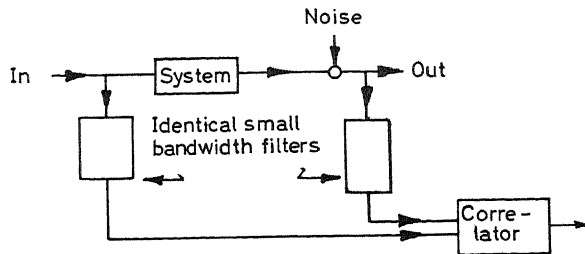


Figure 6. Measurement of the dynamic properties of a system by small bandwidth filtering and correlating

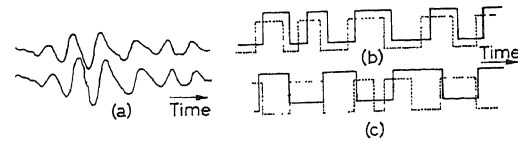


Figure 7. Output signals of the filters in Figure 6. Filter decay is 24 dB/octave to both sides of the centre frequency. (a) Upper trace is filtered input, lower trace is filtered output; (b) clipping after filtering. Dotted curve is filtered input, drawn curve is filtered output; (c) as (b), the filter centre frequency approaches the  $-180^\circ$  phase shift frequency of the system

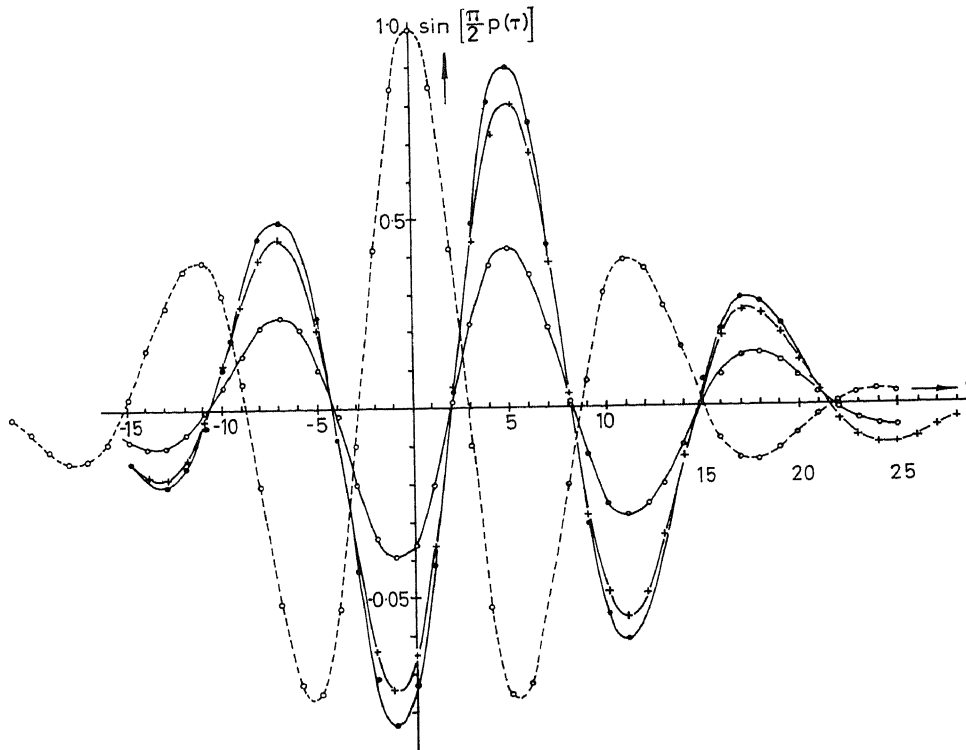


Figure 8. Measured cross-correlation functions  $\sin[\pi/2 \cdot p(\tau)]$  for a third-order system after small bandwidth filtering and clipping of the input and output signal. Dotted line is the auto-correlation function of the filtered input signal. The system consists of three equal time constants of 0.1 sec; filter centre frequency, 2.0 c/sec; sampling distance 44 msec; observation time  $10^5$  samples; input signal 0.5 c/sec; added noise at the output 0.5 c/sec; drawn lines are the cross-correlation for three different signal-to-noise ratios; zero crossings and extrema do not shift along the  $\tau$  axis; the phase shift measured from the shift in zero crossings and extrema around  $\tau = 0$  is  $156^\circ$  (actual phase shift is  $154^\circ$  with the given parameter values)

third-order system for the particular frequency can be measured very accurately, even under the presence of disturbing noise. It is possible in this way to make on-line measurements of the phase characteristic of a process.

A follow-up system for tracking the frequency with  $-180^\circ$  phase shift in a system has been investigated; the procedure shows the rapid responses to changes in the dynamic properties, that are characteristic for model methods. This follow-up system resembles the well-known lock-in detectors that are used in radar techniques for estimating the time delay between an input wave-form and a distorted output wave-form.

#### Appendix I—A Parallel Correlator (Relay Type)

Based on the block diagram in Figure 9 a correlator has been built from commercially available logic modules. The sign of a

signal  $x_1$  is transported in a shift register delay line. Electronic relays can be connected to every shift register stage. These switches pass other signals, provided with the right polarity

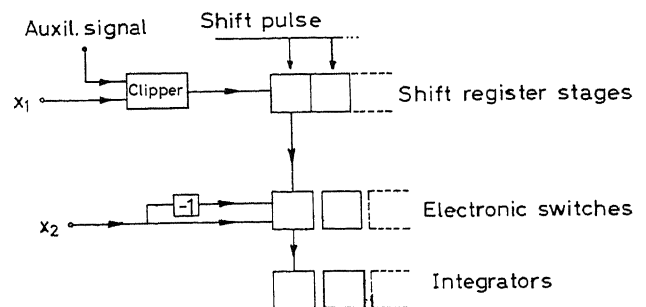


Figure 9. Block diagram of a parallel correlator

depending on the state of the corresponding shift register stage, to integrators. An auxiliary signal can be added at the entrance of the clipping circuit.

The device consists of 50 shift register stages, 25 electronic switches, 3 inverters, 3 clippers with entrances for auxiliary signals and 10 binary dividers for producing sub-frequencies of the shift-pulse frequency (maximum frequency with the used components is 200 kc, allowing a minimum time distance of 5  $\mu$ sec between succeeding points of the correlation function). These units can be interconnected with a patch panel to programme the device for calculation of auto-correlation function, multiple cross-correlation with positive and/or negative time shift, etc. It is evident that this set-up can be simplified to a polarity coincidence correlator by using AND gates with counters instead of electronic relays and integrators.

## Appendix II—A Serial Correlator (Polarity Coincidence Type)

This device has been developed for measuring multiple correlations in slow processes. The polarities of the signal samples are recorded in paper tape according to the diagram in Figure 10.

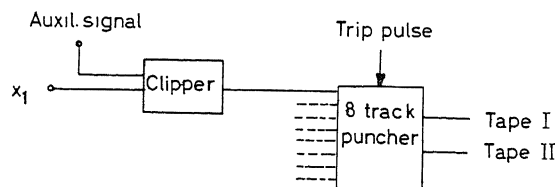


Figure 10. Block diagram of a device for simultaneous recording of the polarity of eight signals with a paper tape puncher with double tape attachment

A positive polarity appears as a hole in a certain track. Eight parallel tracks are possible. Sampling speed ranges from one every 10 sec to 33 samples/sec. Two identical tapes are punched simultaneously. A tape may contain 100,000 samples of a signal in a single track (300 m paper tape).

The reading device consists of a bi-directional punched tape transport mechanism that transports both tapes parallel and

keeps them synchronous with free running, mechanically coupled sprocket wheels. A photo electric reading head gives pulses to a coincidence detector, coincidence of 'no holes' being determined with the sprocket hole. Very high speeds are possible (over 4,000 symbols/sec) as neither constant speed nor immediate braking is necessary. Delay is obtained by shifting the sprocket wheels electromechanically with a ratch wheel construction. This can be done while tape is reversing direction. Counters with an automatic print out display the results. A typical example of the scope of the instrument is: 100,000 samples of four signals (three inputs and one output) are collected in 24 h. Reading off four auto-correlation functions and three cross-correlation functions each consisting of 50 points, is possible within 1 h.

Several hundred correlation functions have been determined with these devices. By and then verification with a digital computer on 2-decimal data have been carried out, with very satisfactory results in all measurements.

## References

- 1 WATTS, D. G. A general theory of amplitude quantization with applications to correlation determination. *Instr. elect. Engrs Monogr.* No. 481 M (1961)
- 2 VELTMAN, B. P. TH. and KWAKERNAAK, H. Theorie und Technik der Polaritätskorrelation für die dynamische Analyse niederfrequenter Signale und Systeme. *Regelungstechnik* No. 9 (1961) 357-364
- 3 VAN VLECK, J. H. The spectrum of clipped noise. *Harv. Radio Res. Lab. Rep.* No. 51 (1943)
- 4 BUSSGANG, J. J. Crosscorrelation functions of amplitude distorted gaussian signals. *M.I.T. Res. Lab. Electron. Tech. Rep.* No. 216 (1952)
- 5 MCFADDEN, J. A. The fourth product moment of infinitely clipped noise. *Trans. Inst. rad. Engrs P.G.I.T.* (1958)
- 6 FULLER, A. T. Sampling errors in the measurement of autocorrelation. *J. Electron. Contr.* 4 (1958) 551-566
- 7 WONHAM, W. M. and FULLER, A. T. Probability densities of the smoothed random telegraph signal. *J. Electron. Contr.* 4 (1958) 567-577
- 8 MCFADDEN, J. A. The correlation function of a sine wave + noise after extreme clipping. *Trans. Inst. rad. Engrs. P.G.I.T.* 2 (1956) 82-83
- 9 LANING, J. H. and BATTIN, R. H. *Random Processes in Automatic Control*. 1956. New York; McGraw Hill

## DISCUSSION

P. JESPERS, *University of Louvain, Belgium*

In the case of multiple input signals, the addition of independent auxiliary functions in the same number as there are input signals provides a correct correlation function for any value of  $\tau$ , whatever the joint probabilities of the input signals may be. Such function generators can be easily realized in analogue form by means of synchronized sawtooth generators, and in digital form by means of quasi-random number generators.

B. P. TH. VELTMAN, and A. VAN DEN BOS, *in reply*

We thank Professor Jaspers for his clarifying remark concerning the addition of the auxiliary signal. To eliminate the influence of the auto-correlation of the auxiliary signals for  $\tau = 0$ , it is indeed advantageous to add different auxiliary signals to the ones that have to be correlated. We disagree, however, that these auxiliary signals are orthogonal.

In fact, the use of deterministic orthogonal auxiliary signals will decrease the additional statistical variance. However, the deterministic orthogonal signals will always have a periodic character in order to be able to generate them easily; this necessitates some care with regard to the sampling period.

In an example of an auto-correlation function of a Gaussian signal with and without auxiliary signals, the variance hardly increases with the additional signal; proper normalization for  $\tau = 0$  will also decrease the variance. The auxiliary signals in this case are simply two triangular waveforms of period  $t_1$  and  $2t_1$ .

J. DELCOUR, *Applied Scientific Research Organization, Lorentzweg, Delft, Netherlands*

In your paper it is stated that the statistical variance of the polarity coincidence procedure is of the same order of magnitude as if the full amplitude range of the signal was taken into account. Can you show any experimental verification of this?

B. P. TH. VELTMAN, and A. VAN DEN BOS, *in reply*

In Figure 2 we show an example of the variances in an auto-correlation function for increasing quantization roughness. These measurements were obtained as follows: with a commercial digital voltmeter, a recording with seven bit accuracy was made on punched tape in a pure binary code from a Gaussian signal. Two identical tapes were punched simultaneously.

With the tape reader described in Appendix II the correlation function was measured with the aid of a special digital multiplier. This multiplier works as follows: a number from tape I is read in a counter; a gate is opened and the counter is emptied by pulses from a 1 Mc generator, then the gate is closed. This procedure is repeated as many times as the number from tape II indicates. The total number of pulses of the 1 Mc generator passing through the presettable counter is measured. By leaving out the less significant bits one is able to calculate the correlation function for increasing quantization roughness.

This way of multiplication and integration proved to be much simpler than binary multiplication and could be realized with relatively few (20) commercial logic circuits.

H. KWAKERNAAK, *University of California, California, U.S.A.*

It is not necessary to make the derivation of section 3(a) as complicated as it is. If  $x_0(t)$  and  $x_n(t)$  are both Gaussian, then  $x_0(t) + x_n(t)$  is also Gaussian.

Therefore the p.c.c.f. of  $x_i(t)$  and  $x_0(t) + x_n(t)$  is given by

$$P_{x_i, x_0+x_n}(\tau) = \frac{2}{\pi} \arcsin R_{x_i, x_0+x_n}(\tau) \quad (1)$$

where  $R_{x_i, x_0+x_n}(\tau)$  is the normalized cross-correlation function. If  $x_i(t)$  and  $x_n(t)$  are uncorrelated, this normalized cross-correlation function can easily be found to be

$$R_{x_i, x_0+x_n}(\tau) = \frac{K_{x_i, x_0}(\tau)}{\sqrt{K_{x_i, x_i}(0)[K_{x_0, x_0}(0) + K_{x_n, x_n}(0)]}} \quad (2)$$

where  $K_{x_i, x_0}(\tau)$  represents the (not normalized) cross-correlation function of  $x_i(t)$  and  $x_0(t)$ , etc. It is simple to find that (a) can be put in the form

$$P_{x_i, x_0+x_n}(\tau) = \frac{2}{\pi} \arcsin \frac{R_{x_i, x_0}(\tau)}{\sqrt{1 + \frac{K_{x_n, x_n}(0)}{K_{x_0, x_0}(0)}}}$$

Clearly, by determining

$$\sin \frac{\pi}{2} \cdot P_{x_i, x_0+x_n}(\tau)$$

one finds a function which is proportional to the desired cross-correlation function  $R_{x_i, x_0}(\tau)$ .

The proportionality factor depends upon the signal-to-noise ratio

$$\frac{K_{x_0, x_0}(0)}{K_{x_n, x_n}(0)}$$

It can be seen quite clearly from Figure 3 that this phenomenon indeed occurs.

B. P. TH. VELTMAN, and A. VAN DEN BOS, *in reply*

Dr. Kwakernaak's remark is much appreciated. In the case of Gaussian disturbances a direct proof of the applicability of the polarity method with cross-correlation can indeed be given. Although we were aware of this, we never verified the result in that case. It is very interesting to see that the correlation function will only be multiplied by a constant factor.

This gives certain possibilities for estimating the power of the noise compared to the signal power. A situation occurs accordingly with the addition of auxiliary signals, where the known power of the auxiliary signal admits an estimation of the signal power.

# Notes sur une Fonction Aléatoire d'Aspect Physique

M. J. PÉLÉGRIN

## Summary

The study of systems increasingly requires the use of random functions to represent both inputs and noise. In the automatic control field, the use of statistical criteria makes possible the computation of the economic efficiency of the system when working under conditions close to real operation.

This paper attempts to define a process generating a random function having some properties which give it a 'physical aspect': limited maximum acceleration, limited amplitudes, or still better, probability density of given amplitudes.

Preliminary tests lead to definite functions with constantly maximum acceleration (positive or negative): the construction of such functions does not present difficulties and one can use three parameters to fit the curve to the physical phenomena to be simulated; however, it is not possible to impose a probability density in amplitude.

Finally, a generating process is given for the imposed maximum acceleration, the probability density (curve) and at least a parameter which is used to adjust the time scale. The function is made of a succession of parabolic arcs with a common tangent at their points of junction. Samples have been built as well as the corresponding autocorrelation function. The direct computation of some mathematical functions (particularly the autocorrelation function) attached to these random functions is being studied.

## Sommaire

L'étude des systèmes implique de plus en plus l'usage de fonctions aléatoires pour représenter soit les entrées, soit les bruits. En Automatique Industrielle, l'usage de critères statistiques permet de calculer le rendement économique du système dans des conditions proches du fonctionnement réel.

Cette note a pour but de définir un processus générateur de fonction aléatoire ayant certaines propriétés qui lui confèrent un « aspect physique »: accélération maximale bornée, amplitudes bornées, ou mieux, densité de probabilité en amplitude imposée.

Des essais préliminaires ont conduit à définir des fonctions à accélération toujours maximale (en module): la construction est simple, on dispose de 3 paramètres pour « ajuster » la courbe mais on ne peut pas imposer *a priori* une densité de probabilité en amplitude.

On donne alors un processus générateur pour une accélération maximale imposée, une densité de probabilité imposée (courbe) et au moins un paramètre pour ajuster l'échelle des temps. La fonction est constituée par une succession d'arcs de parabole se raccordant tangentielllement. Des échantillons ont été construits ainsi que la fonction d'autocorrélation correspondante. Le calcul direct de certaines fonctions mathématiques, la fonction d'autocorrélation en particulier, attachées à ces fonctions aléatoires est en cours d'étude.

## Zusammenfassung

Die Systemuntersuchungen erfordern mehr und mehr den Gebrauch von Zufallsfunktionen zur Darstellung sowohl der Eingangs- als auch der Rauschsignale. In der industriellen Automatisierung gestattet die Verwendung statistischer Kriterien die Berechnung der Wirtschaftlichkeit eines Systems unter wirklichkeitsnahen Bedingungen.

Die Arbeit macht den Versuch, ein Verfahren zur Erzeugung einer Zufallsfunktion zu definieren, die den physikalischen Bedingungen möglichst nahekommt: eine beschränkte maximale Beschleunigung, begrenzte Amplituden oder besser eine Wahrscheinlichkeitsverteilung der gegebenen Amplituden.

Vorläufige Untersuchungen führten zu Funktionen mit stets maximaler positiver oder negativer Beschleunigung. Die Erzeugung einer solchen Funktion ist einfach, man verfügt über 3 Parameter, um die Kurve der simulierten physikalischen Erscheinung anzupassen, aber es war nicht möglich, von vornherein eine Wahrscheinlichkeitsverteilung der Amplitude vorzugeben.

Es wird daher ein Verfahren zur Erzeugung von Funktionen bei vorgegebener Größe der maximalen Beschleunigung und der Wahrscheinlichkeitsverteilung der Amplitude sowie eines Parameters zur Anpassung des Zeitmaßstabes, vorgeschlagen. Die Funktion wird durch eine Folge von Parabelbögen, die sich tangential aneinander anschließen, gebildet. Beispiele wurden aufgestellt, sowie auch die entsprechende Autokorrelationsfunktion. Die direkte Berechnung gewisser mathematischer Ausdrücke auf Grund dieser Zufallsfunktionen wird noch untersucht.

## Objet

L'objet de cette note est de définir un processus générateur d'une fonction aléatoire satisfaisant les conditions suivantes: (a) densité de probabilité des amplitudes données; (b) accélération (dérivée seconde) maximale donnée en module; c'est ce caractère qui confère principalement l'aspect physique de la fonction cherchée; (c) possibilité d'ajustement de l'échelle des temps (il s'agit d'adapter au mieux le spectre de la fonction cherchée); (d) principales propriétés statistiques calculables (fonction d'autocorrélation, spectre, etc...), et (e) échantillons facilement reproductibles.

## Intérêt de ces fonctions

Habituellement, l'étude des systèmes se fait à partir de sollicitations sinusoïdales. Si le système étudié est linéaire il suffit d'étudier le rapport des amplitudes de l'entrée et de la sortie, et le déphasage entre l'entrée et la sortie, pour toutes les fréquences du domaine à étudier. Mais si le système n'est pas linéaire, pour chaque fréquence, il faut étudier la réponse à diverses amplitudes. Bref, il faut explorer un domaine du plan fréquences/amplitudes et de ce fait l'étude du système est beaucoup plus complexe. L'étude expérimentale des réponses à des sollicitations sinusoïdales nécessite la stabilisation du système pour chaque fréquence et chaque amplitude et devient très longue; d'ailleurs on tend de plus en plus à étudier non pas la réponse du système à une sollicitation (fréquence amplitude) mais la réponse du système à une entrée non permanente en utilisant des critères statistiques (moyennes). C'est en particulier le cas de tous les critères fondés sur l'erreur quadratique moyenne minimale.

On sait alors que l'erreur moyenne dépend de la fonction d'entrée. L'idée de cette note est de définir des fonctions susceptibles d'être injectées dans le système de façon aussi proche que possible des fonctions réelles auxquelles le système devra répondre, et de calculer l'erreur quadratique moyenne à partir de ces

fonctions. Les fonctions qui sont définies ci-dessous peuvent être ajustées, dans une certaine mesure, à des phénomènes réels. En particulier l'accélération maximale d'un signal réel est toujours connue (ou tout au moins une borne supérieure, qu'on peut en général déterminer sans trop de surabondance). C'est l'un des paramètres principaux dont on dispose dans la définition du processus générateur. Il est évident que si les propriétés statistiques de cette fonction aléatoire peuvent être connues mathématiquement, *a priori*, l'ajustement de la fonction aléatoire au phénomène réel sera d'autant plus simple.

On dispose donc d'un processus générateur permettant de définir une fonction aléatoire dont les propriétés mathématiques sont connues; l'étude du système peut se faire en injectant à l'entrée une telle fonction (puisqu'on la sait suffisamment proche de la fonction réelle); on peut mesurer ou calculer l'erreur quadratique moyenne entre l'entrée et la sortie du système.

### Conditions à satisfaire

La condition principale concerne l'accélération de la courbe. Afin de pouvoir calculer les principales propriétés de la fonction on définit d'abord une fonction en créneaux dont les instants de sauts suivent une loi de Poisson. Il est évident qu'on peut prendre toute autre loi pour déterminer ces sauts; cependant la loi de Poisson s'adapte d'emblée relativement bien à un grand nombre de phénomènes physiques.

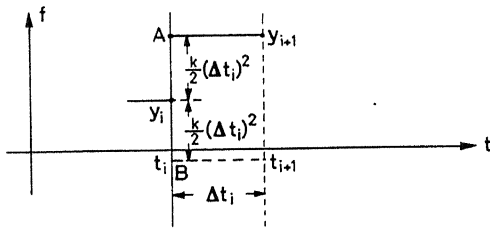


Figure 1

La condition d'accélération maximale sera définie comme suit. Soit  $t_i$  un instant auquel un saut se produit; l'amplitude du saut sera proportionnelle au carré de l'intervalle  $\Delta t_i = t_{i+1} - t_i$ . Si  $k$  est l'accélération maximale imposée, le coefficient de proportionnalité sera égal à  $k/2$  (Figure 1).

Cette condition implique que la succession des instants  $t_i$  soit connue au moins un pas en avance. Cette condition ne

présente pratiquement pas de difficultés (on notera cependant que l'intervalle séparant 2 instants peut, dans une loi de Poisson, atteindre n'importe quelle valeur, ce qui nécessiterait un dispositif à mémoire à durée illimitée). Mais il s'agit ici d'un problème physique et, si l'on admet qu'une probabilité de  $10^{-8}$  par exemple, est négligeable, il suffira de prendre comme durée maximale de la mémoire, la durée de l'intervalle  $\Delta t_i$  correspondant à cette probabilité; l'utilisation d'une loi telle que la loi de Poisson rend aisée cette détermination.

### Essais préliminaires

#### Fonctions à accélération toujours maximale

Dans les premiers essais que nous avons fait on s'imposait à chaque saut une accélération maximale, mais le signe de cette accélération était tiré au sort. La fonction primaire en créneaux correspondante tend à diverger en amplitude (résultat classique, l'écart quadratique moyen est proportionnel à la racine carrée du nombre de sauts). Il est possible de ramener la fonction vers l'axe, c'est-à-dire, de diminuer le coefficient de proportionnalité, en imposant un tirage au sort du signe de l'accélération non équiprobable. Par exemple, saut négatif avec une probabilité de 60 pour cent si la dernière valeur en amplitude est positive. Des calculs d'échantillons sur machines à calculer ont montré qu'effectivement la courbe se stabilisait de façon satisfaisante. La Figure 2 représente un échantillon répondant aux conditions suivantes: accélération toujours maximale en module; tirage 40—60 pour cent — la courbe en pointillé correspond à un tirage équiprobable. Un tel processus a l'avantage de la simplicité mais il est difficile de connaître les propriétés mathématiques statistiques de cette fonction aléatoire. La détermination de la fonction de distribution conduit à des calculs inextricables... On peut cependant ajuster dans une certaine mesure les paramètres du processus dans le cas où les instants  $t_i$  satisfont une distribution de Poisson, on dispose:

(a) du nombre moyen  $a$  de points par unité de temps (seconde) — alors la probabilité d'avoir  $n$  points dans un intervalle de temps  $\Delta t$  est donnée par:

$$p(n, \Delta t) = \frac{e^{-a\Delta t} (a\Delta t)^n}{n!}$$

(b) du paramètre de proportionnalité  $k$  définissant à la fois l'amplitude du saut et l'accélération maximale imposée — si la fonction représente des volts  $k$  s'exprime en volt/s<sup>2</sup>.

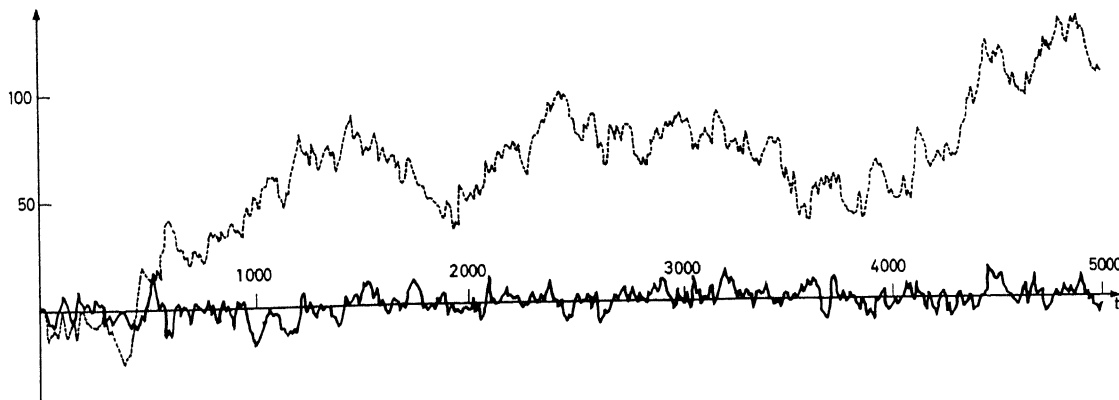


Figure 2

(c) la probabilité  $p$  définissant le rappel vers l'axe, et par conséquent 'l'échelle' de la stationarité et la dispersion des amplitudes de la fonction.

Le paramètre  $a$  est fortement lié à la notion de fréquences contenues dans l'épreuve de la fonction; le paramètre  $k$ , une fois fixé  $a$ , définit le domaine de variation (en probabilité) de la fonction; le paramètre  $p$  permet d'obtenir la stationarité plus ou moins rapidement et fixe la dispersion désirée des amplitudes. Bien entendu, on peut utiliser toute loi de distribution des instants  $t_i$ , y compris des lois empiriques.

#### Fonctions à accélérations bornées mais non nécessairement toujours maximales

(a) Si, au lieu de tirer au sort l'un des deux points  $y_i \pm k/2 (\Delta t_i)^2$  on fait un tirage, suivant une certaine loi, sur l'intervalle  $k (\Delta t_i)^2$  centré sur  $y_i$  on obtiendra une fonction à accélération quelconque mais bornée. La Figure 3 représente un échantillon pour lequel on ferait un tirage équiprobable sur  $AB$ .

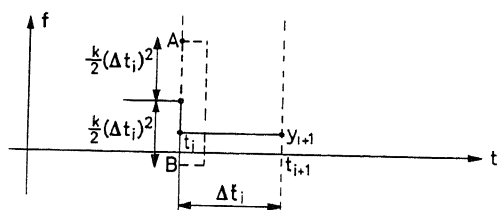


Figure 3

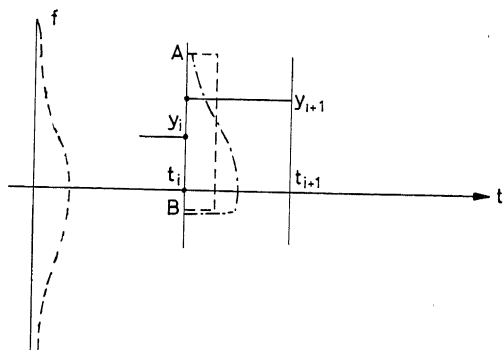


Figure 4

Ce processus n'empêche pas la divergence en valeur quadratique moyenne déjà signalée. On peut alors *stabiliser* la courbe autour de l'axe, par exemple de la façon suivante: on pondère la loi de probabilité centrée sur la valeur à  $t_i$  (cf. Figure 4; on suppose toujours que le tirage est équiprobable sur  $AB$ ) par une loi de probabilité centrée sur l'axe des temps (par exemple une loi gaussienne comme indiqué sur la Figure 4). Le tirage au sort s'effectue alors suivant une loi de probabilité, variable à chaque saut (courbe en trait mixte).

Une variante, plus brutale, consiste à prendre pour loi fixe, centrée sur l'axe des  $t$  une densité constante de  $-M$  à  $+M$ ,  $M$  étant la borne d'amplitude imposée pour la fonction. Le processus est alors très simple, puisqu'il consiste à remplacer le tirage au sort équiprobable sur  $AB$  par un tirage équiprobable sur  $A'B$  (Figure 5).

La Figure 6 représente un échantillon de 100 sauts environ, lissé, obtenu sur machine CAB 500 répondant aux conditions

suivantes: les changements d'états de la fonction en créneaux suivent une loi de Poisson (nombre moyen de changements par unité de temps: 1) bornes d'amplitude;  $M = \pm 5$ . L'inconvénient de tous ces processus est de ne pas permettre l'ajustement de la densité de probabilité des ordonnées; le calcul de cette densité est très difficile sinon impossible comme déjà signalé. Nous proposons donc le processus générateur suivant.

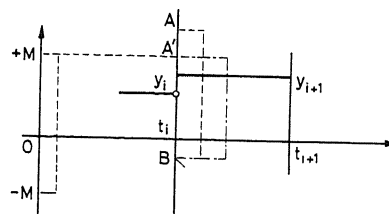


Figure 5

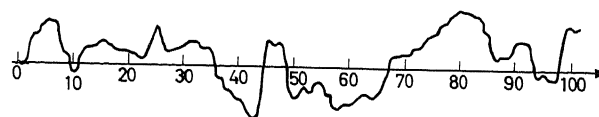


Figure 6

#### Processus proposé

##### Généralités sur le processus cherché

Comme indiqué au premier paragraphe, nous cherchons une fonction aléatoire ayant une accélération maximale bornée en module. Nous n'imposons cependant pas que la fonction soit toujours à accélération maximale; l'accélération à chaque instant peut avoir une valeur comprise entre les 2 bornes (positive et négative). On verra plus loin que cette condition permet d'ajuster la fonction sur une loi de probabilité en amplitude donnée.

##### Fonction génératrice primaire (fonction en créneaux)

On définit d'abord une fonction en créneaux, les instants de saut étant définis, par exemple, par une loi de Poisson. On admet que la densité de probabilité en amplitude est imposée; soit  $G(Y)$  cette fonction (Figure 7). Supposons qu'à l'instant considéré l'échantillon de la fonction aléatoire soit tel qu'il satisfasse à une loi de densité de probabilité en amplitude constante: soit  $F_0(Y)$  cette loi. La différence  $G(Y) - F_0(Y)$  est certainement positive dans un ou plusieurs intervalles de  $(Y)$ :  $(Y_1, Y_2)$  et  $(Y_3, Y_4)$  sur la Figure 7(b).

Soit  $t_i$  un instant auquel un saut doit se produire et immédiatement après que l'échantillon calculé jusqu'à cet instant ait pour densité de probabilité la densité constante  $F_0(Y)$ . Nous voulons maintenant que la suite de l'échantillon tende vers une loi de probabilité  $G(Y)$ . A l'instant  $t_i$  nous ouvrons une première porte (voir Figure 7(a)) d'amplitude totale  $AB = k(t_{i+1} - t_i)^2$  centrée sur la dernière ordonnée à  $(t_i - 0)$ ;  $k$  désigne l'accélération maximale imposée (module).

On cherche l'intersection de cette porte et des intervalles où la différence  $G(Y) - F_0(Y)$  est positive. On fait un tirage au sort, par exemple équiprobable, dans les intervalles communs à ces deux domaines (intervalles  $(B, Y_2)$  et  $(Y_3, A)$  sur la Figure 7). On obtient ainsi la valeur du créneau suivant.

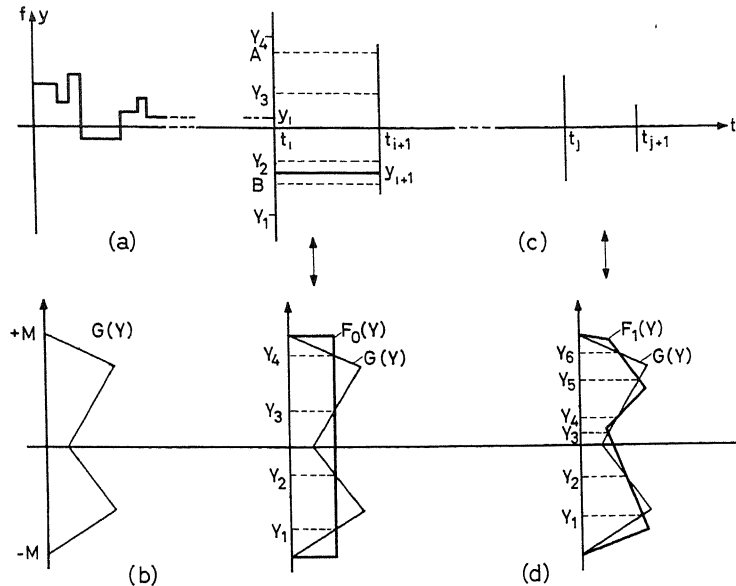


Figure 7

On continue cette méthode pendant un certain nombre de sauts (par exemple 1000 sauts) puis on calcule la densité de probabilité en amplitude de l'échantillon ainsi construit, soit  $F_1(Y)$  cet échantillon. Par rapport à la fonction initiale  $F_0(Y)$ ,  $F_1(Y)$  s'est rapproché de  $G(Y)$  (Figure 7 (d)).

On recommence le même processus que ci-dessus mais en effectuant maintenant la soustraction  $G(Y) - F_1(Y)$  et en ne retenant que les intervalles où cette différence est positive. On garde ce ou ces intervalles (il y en a 3 sur la Figure 7(d)) durant un prochain lot de sauts (1000 par exemple). Après quoi on calcule la nouvelle loi de probabilité  $F_2(Y)$  qu'on gardera durant le prochain lot de sauts.

On peut ainsi espérer tendre après un certain nombre de lots de sauts, vers un processus stationnaire. Il est évident, en effet, que les premiers sous-échantillons ne sont pas stationnaires puisque, par leur construction même, ils évoluent lentement en densité de probabilité vers la loi imposée  $G(Y)$ .

Lorsque la fonction  $F_n(Y)$  est très voisine de  $G(Y)$  on notera qu'il existe de toutes façons des intervalles où  $G(Y) - F_n(Y)$  est positif. Cependant, afin d'obtenir une stationnarité plus fine, il peut être avantageux de réduire la période de mise à jour de la fonction de densité de probabilité en amplitude (par exemple 100 sauts). Ce point ne constitue pas un obstacle à ce que nous recherchons: il s'agit d'une fonction aléatoire stationnaire d'aspect physique.

La stationnarité ne saurait être obtenue dans un intervalle au moins égal à une certaine durée: réduire cet intervalle reviendrait à restreindre le caractère aléatoire de la fonction.

La fonction en créneaux doit être lissée par des arcs de paraboles (Figure 8). Chaque saut est remplacé par deux arcs de paraboles égaux mais de concavité opposée. Ils se raccordent au milieu de l'intervalle  $t_{i+1} - t_i$ . L'accélération reste donc constante de  $t_i$  à  $(t_{i+1} - t_i)/2$  et prend une valeur opposée dans le demi intervalle suivant. Un tel lissage ne permet pas de passer simplement des propriétés statistiques de la fonction en créneaux à celles de la fonction lissée, mais par contre on est assuré que

l'accélération de la fonction ne dépassera jamais la borne maximale imposée à quelque instant qu'on se place.

On peut cependant dans le cas présent calculer mathématiquement la densité de probabilité des amplitudes de la nouvelle fonction.

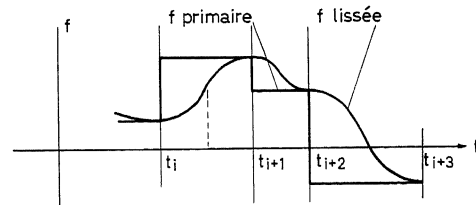


Figure 8

## Résultats

### Données

La mise au point du processus a été faite sur une machine SEA-CAB 500 de l'ENSA. Le calcul de la fonction définitive a été exécuté par le laboratoire de calcul de l'institut Polytechnique de Grenoble\* sur une machine BULL-Gamma III E.T. Le processus utilisé est celui décrit ci-dessus utilisant une suite poissonnienne de points en abscisse. (Les propriétés statistiques de cette suite ont été vérifiées.)

La courbe de densité d'amplitudes a été remise à jour tous les 250 points et elle a été imprimée tous les 1000 points. Comme indiqué au § 5 on suppose, au début du calcul, que la densité de probabilité de la 'fonction passée' est uniforme. On fait le tirage au sort (équiprobable) dans l'intervalle intersection de (a) l'intervalle  $y_i \pm (\Delta t_i/2)^2$ , (b) l'intervalle  $\pm A$ ,  $A = 5$  (les amplitudes ont été quantifiées en 32 niveaux), et (c) les intervalles où la

\* Nous remercions Mr le Directeur du Service Technique Aéronautique qui a bien voulu supporter les frais de ces calculs, Mr le Professeur Kuntzmann et Mr Belino, respectivement Directeur et Assistant du Laboratoire de Calcul de Grenoble pour leur collaboration à cette étude.

différence: densité de probabilité imposée moins densité de probabilité actuelle est négative.

Pour ce tirage au sort on a utilisé une table de nombres au hasard équiprobables (propriétés statistiques vérifiées). La densité de probabilité évolue vers la courbe de densité imposée (Figure 9).

Les 5 000 derniers sauts pour lesquels la courbe de densité est très voisine de la courbe imposée, ont été sortis sur cartes puis lissés par des arcs de paraboles (pas 0,25). La courbe finale a été sortie sur cartes afin d'en calculer la fonction d'autocorrélation.

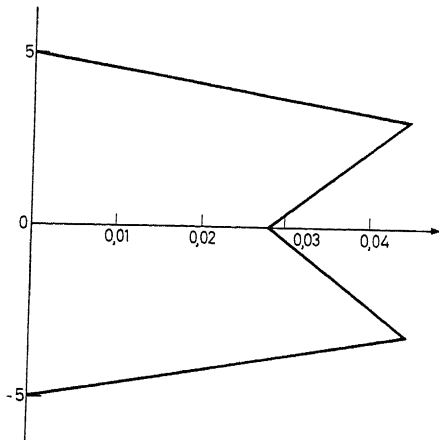


Figure 9

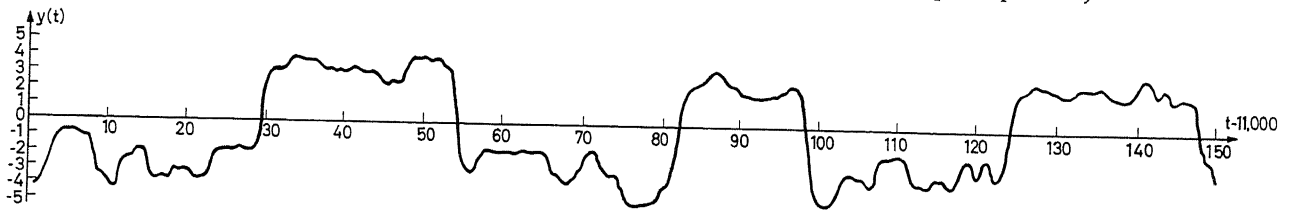
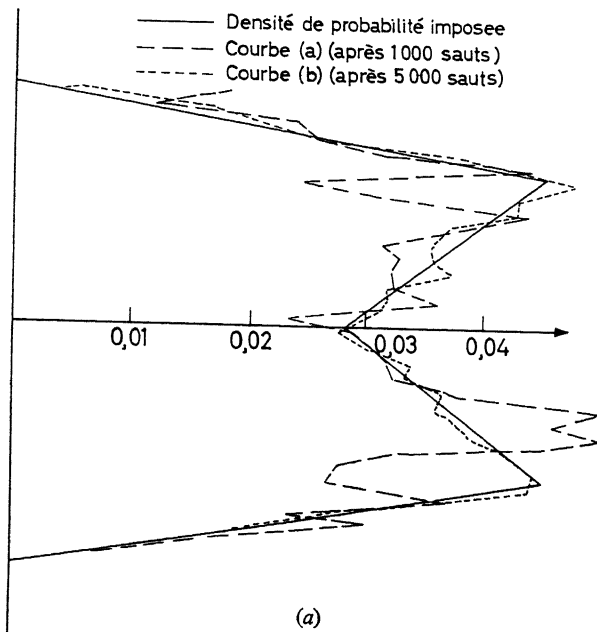
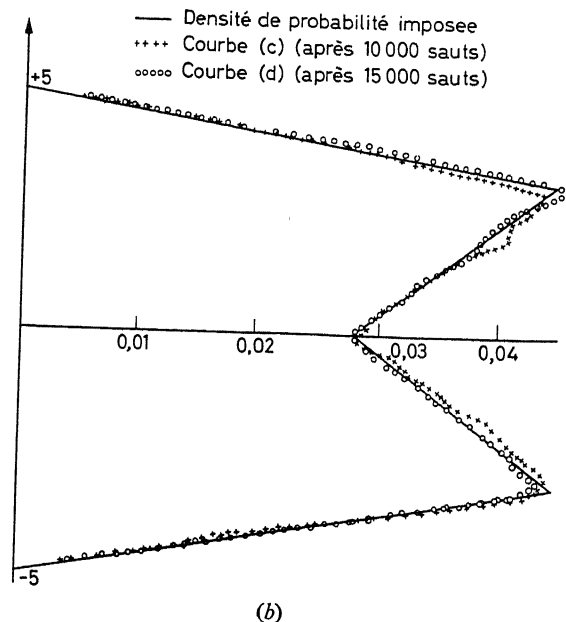


Figure 10



(a)



(b)

Figure 11

### Résultats numériques

La Figure 10 montre la partie de la fonction finale (lissée) comprise entre les points  $t = 11\,000$  et  $t = 11\,150$  (lissage au pas de 0,25). Les Figure 11(a) et 11(b) montrent l'évolution de la courbe de densité de probabilité des amplitudes: courbe (a) après 1 000 sauts; courbe (b) après 5 000 sauts; courbe (c) après 10 000 sauts, et courbe (d) après 15 000 sauts.

La Figure 12 représente la fonction d'autocorrélation portant sur un échantillon de 4 000 points de la fonction lissée; l'échantillon utilisé a été prélevé dans les 5 000 derniers sauts calculés avant lissage soit 20 000 points après lissage. Pour le calcul de la fonction d'autocorrélation on a utilisé des sous-échantillons de 4 000 points. La Figure 12 correspond au sous-échantillon compris entre les 4 000<sup>e</sup> et 8 000<sup>e</sup> points.

### Calcul direct des propriétés statistique

Puisque la densité de probabilité constitue une donnée du problème, la propriété statistique la plus intéressante à déterminer est la fonction d'autocorrélation ou le spectre. Nous avons abordé ces calculs mais ils sont inextricables dans toute leur généralité. Aussi avons-nous recherché une expression approchée de la fonction d'autocorrélation en procédant ainsi.

L'approximation consiste à ne calculer que les termes apportant une contribution importante dans la fonction d'autocorrélation.

† On appelle 'saut' les instants où la fonction en créneaux subit une discontinuité et 'points' les abscisses ou l'ordonnée de la fonction lissée a été calculée. (Il y a 4 points par saut.)



tion: ce sont les termes correspondant à de petits décalages. On note, en effet, sur la *Figure 12* que la corrélation tombe en dessous de 0,15 pour un décalage inférieur à 20 sauts. On envisage alors les différents cas possibles: pas de saut dans l'intervalle  $\tau$ , un saut, deux sauts, etc. ..., on calcule les corrélations dans chacun

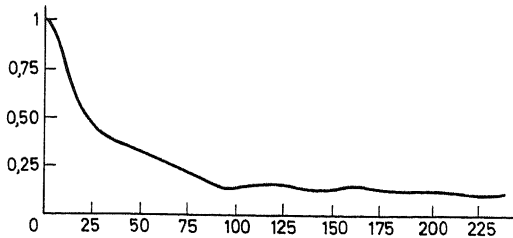


Figure 12

des cas et on les affecte de la probabilité de réalisation de chacun de ces cas. Ces probabilités sont connues et simples dans le cas d'une loi de Poisson. La valeur de  $\tau$  est elle-même liée à l'échelle des temps adoptée: si on a pris  $\lambda$  sauts, en moyenne, par seconde la durée moyenne d'un saut est  $1/\lambda$  seconde et  $\tau$  doit être surtout étudié dans l'intervalle  $0,05 \ 1/\lambda$  et  $5/\lambda$ . Alors le nombre

de termes intervenant dans la fonction d'autocorrélation est limité. L'expression complète est cependant compliquée.

### Conclusions

Le processus générateur de la fonction aléatoire suggéré permet d'obtenir des échantillons ayant les propriétés suivantes. (a) l'accélération est bornée en module — l'accélération peut être toujours maximale ou peut varier entre les deux bornes (positive et négative); (b) la densité de probabilité des amplitudes est imposée; (c) l'échelle des temps est ajustable par la loi de probabilité de la fonction génératrice primaire (créniaux). Si on utilise une loi de Poisson on dispose du nombre moyen de sauts par unité de temps. (d) elle est facilement reproductible par machine à calculer même en 'temps réel'.

L'utilisation de ces fonctions semble particulièrement indiquée en Automatique pour la recherche d'optimums soit par voie expérimentale (on injecte une telle fonction et on effectue le réglage des paramètres avant la mise en service du système) soit par voie semi expérimentale. Une fois les principales propriétés de la fonction estimées (amplitude, accélération, échelle des temps, densité de probabilité) on peut calculer un échantillon puis sa fonction d'autocorrélation, puis son spectre et accessoirement d'autre propriétés statistiques.

### DISCUSSION

F. MESCH, *Institut für Regelungstechnik, TH Darmstadt, Germany*

At the beginning of the paper it is stated that the object was to generate a process with adjustable amplitude density, adjustable maximum acceleration and adjustable power spectrum. I would like to ask how far it was possible to adjust the amplitude density and the power spectrum independently, since the relationship between these is known to be very involved.

M. J. PÉLÉGRIN, *in reply*

The adjustment I referred to concerned a limit of the acceleration but not the acceleration spectrum itself. This one is directly related to the amplitude spectrum; this is the second derivative of the autocorrelation function or the product by  $(j\omega)^2$  of the amplitude spectrum for a stationary random process. Limitations of the modulus of the acceleration give very loose restrictions on the amplitude spectrum (particularly in the high frequencies). However, this compatibility is the result of the physical aspect that we are looking for.

# An Uncertainty Relation for Linear Mathematical Models

B. QVARNSTRÖM

## Summary

A relation is derived by means of which the uncertainty of an arbitrary parameter pertaining to a linear mathematical model can be calculated. It is assumed that the model is identified with the process, industrial or other, on the basis of finite time terminal data records, according to a least mean square error criterion. The process is assumed to contain stationary noise sources distorting the output record.

The uncertainty relation is valid for deterministic and stochastic input signals to the process and for all measuring and evaluation procedures that satisfy the least mean square error criterion. Well-known procedures of that kind are the conventional frequency response method and the transfer function determination by the Wiener-Hopf integral equation on the basis of statistical data.

The parameter error R.M.S. value, or the parameter uncertainty, is related to the following factors: the process input signal and the process noise, represented by their power density spectra, the time period of measuring and a parameter influence function.

The information disclosed by the uncertainty relation is dealt with and two examples given.

## Sommaire

On a obtenu une relation permettant de calculer l'incertitude d'un paramètre arbitraire dans un modèle mathématique linéaire. Ce dernier, représente un processus, industriel ou non, à partir des données enregistrées durant un intervalle de temps fini, et dépouillées selon le critère des moindres carrés. On suppose que le processus contient des sources de bruit stationnaires déformant le signal de sortie enregistré.

La relation obtenue est valable pour des signaux déterministes et stochastiques et pour toutes les procédures de mesure ou d'évaluation utilisant le critère des moindres carrés. Parmi ces procédures, se trouvent la méthode fréquentielle classique et celle de détermination des fonctions de transfert à l'aide de l'équation intégrale de Wiener-Hopf sur la base de données statistiques.

Les facteurs influençant la valeur de l'incertitude sont: le signal d'entrée et le bruit du processus, représentés par leur densité spectrale, la durée de la mesure et une fonction d'influence paramétrique.

Le contenu informationnel de la relation est discuté. Deux exemples sont donnés.

## Zusammenfassung

Mit Hilfe der hier abgeleiteten Beziehung kann man die Unbestimmtheit eines beliebigen Parameters in einem linearen mathematischen Modell berechnen. Es wird angenommen, daß die Registrierung der Ein- und Ausgangsgröße über ein endliches Zeitintervall die Annahme des Modells für den (industriellen oder anderen) Prozeß rechtfertigt, wobei als Kriterium der kleinste quadratische Mittelwert des Fehlers zugrunde liegt. Der Prozeß enthalte Quellen für stationäres Geräusch, das die Registrierung der Ausgangsgröße verfälscht.

Die Unbestimmtheitsrelation gilt für deterministische und stochastische Eingangsgrößen und für alle Meß- und Auswerteverfahren, die dem Kriterium des kleinsten quadratischen Mittelwertes des Fehlers genügen. Wohlbekannte Verfahren dieser Art sind das übliche Frequenzgangverfahren und die Bestimmung der Übertragungsfunktion durch die Wiener-Hopfsche Integralgleichung auf Grund statistischer Daten.

Die Unbestimmtheit des Parameters (Effektivwert des Fehlers) steht mit den folgenden Größen in Beziehung: dem Eingangssignal

und dem Rauschen des Prozesses, die durch das Leistungsspektrum dargestellt sind, der Meßzeit und einer Parametereinflußfunktion.

Die durch die Unbestimmtheitsrelation gewonnene Erkenntnis wird erläutert; zwei Beispiele sind angeführt.

## Introduction

The background of the work to be presented is the problem of identifying an industrial or other process, and a mathematical model, in order to prepare for the design or the adjustment of a control system regulating the process. The information available about the process transfer characteristics is assumed to be a finite time record of the process input and output variables. It is further assumed that the output is distorted by a noise component, added by sources within the process or in the measuring device.

A linear mathematical model is used to represent the process, which therefore should be at least approximately linear. The model is specified by a number of parameters; these are to be adjusted so that the model, subject to the actual input signal, will produce an output as similar as possible to the real process output. A least mean square error criterion is used here—this implies that the model shall not contain noise sources, because such sources, if uncorrelated to the corresponding sources of the real process, will on average only increase the mean square value of the output difference.

The best possible model is obtained after a proper adjustment of the model parameters. This model will not be exact, however, because the process noise and the finite observation time will make the parameter values uncertain, to some degree.

In this paper the uncertainty of an arbitrary parameter is estimated and general expressions given, relating the parameter uncertainty to important factors of the model design problem. Those expressions may be used to answer questions such as: for how long is it necessary to measure to allow a specified model accuracy? How complex and detailed can the model be made on the basis of a given amount of experimental data? What kind of input signal will give the least uncertainty of the parameter values?

Regarding the nature of the input signal, no specific assumption is necessary at this stage. Being recorded and thus completely known, it is of no importance if the input signal was generated by a deterministic or a stochastic source. It is shown later that a record of the input signal is not needed in the stochastic case, provided the statistical properties of the signal source are known.

The first step towards the parameter uncertainty is a computation of the parameter deviation in a single test, under the auxiliary assumption that the process noise component is explicitly known. In the second step, the average square value of

the said deviation is derived for a large number of tests. In this procedure the explicit noise signal is replaced by a statistical function characterizing the noise source. The R.M.S. value of the parameter deviation is considered as a measure of the parameter uncertainty.

### The Parameter Deviation in a Single Test

The difference  $e(t)$  between the process and the model output signals is by definition:

$$e(t) = y(t) + n(t) - q(t) \quad (1)$$

The mean square value of the difference can be written as follows:

$$\overline{e^2} = \overline{y^2} + \overline{n^2} + \overline{q^2} + 2\overline{yn} - 2\overline{yq} - 2\overline{nq} \quad (2)$$

The bar indicates the mean value over the period of time during which the signals have been recorded.

The model is adjusted by means of  $N$  parameters  $a_i$  in such a way that the mean square value of eqn (2) will arrive at a minimum. The signal  $q(t)$  is also a function of the parameters  $a_i$ , and the differentiation with respect to these leads to the following  $N$  equations:

$$\frac{d\overline{e^2}}{da_i} = 2\overline{q \frac{\partial q}{\partial a_i}} - 2\overline{y \frac{\partial q}{\partial a_i}} - 2\overline{n \frac{\partial q}{\partial a_i}} = 0 \quad (3)$$

or

$$(q - y) \frac{\partial q}{\partial a_i} = n \frac{\partial q}{\partial a_i} \quad (4)$$

It is now assumed that the set of parameters is chosen in such a way that the model and the process are identical for a certain set of parameter values, denoted by  $a_{i0}$ . Due to the process noise and the limited time of measurement, the parameter values actually obtained in the minimizing procedure will deviate from the values  $a_{i0}$ . At the same time, the measurement, if serving any purpose, has to be so good that the deviations are small. A Taylor series expansion is then possible.

$$q - y = \sum_{k=1}^{\infty} \frac{1}{k!} \left[ \sum_{j=1}^N da_j \left( \frac{\partial}{\partial a_j} \right)_{a_{j0}} \right]^k q \quad (5)$$

Under the assumption that terms of an order higher than one can be neglected, eqn (5) is introduced into eqn (4).

$$\sum_{j=1}^N da_j \overline{\left( \frac{\partial q}{\partial a_i} \right)_{a_i} \left( \frac{\partial q}{\partial a_j} \right)_{a_{j0}}} = n \overline{\left( \frac{\partial q}{\partial a_i} \right)_{a_i}} \quad (6)$$

The parameter deviation of interest may be obtained by the solution of the  $N$  linear, simultaneous eqns (6). Further restrictions will, however, be introduced with respect to the set of parameters in order to make possible the desired result. That is, a formula, valid for a single arbitrary parameter of interest, relating the parameter deviation to the most important factors of the problem; such as the necessary time of recording, the nature of the parameter and so on. Eqn (6) requires that a complete set of parameters must be defined before the probable deviation of any single parameter can be estimated; this is because there is generally a coupling effect between the different parameters.

In the following, only one parameter will be considered—that

is the interesting one, denoted by  $a_i$ . The choice of this parameter will be unrestricted. The other parameters, involved in the same minimizing procedure, will be defined to fulfil the following conditions of orthogonality.

$$\overline{\left( \frac{\partial q}{\partial a_i} \right)_{a_i} \left( \frac{\partial q}{\partial a_j} \right)_{a_{j0}}} = 0 \quad j \neq i \quad (7)$$

The orthogonal parameters can be constructed as linear combinations of the original parameters plus extra, auxiliary parameters. The volume of the computation will increase by the rearrangement of the parameters and by the fact that the minimizing procedure has to be repeated for every parameter investigated. This is of no importance, because the purpose here is to estimate the parameter errors only, and it is not necessary actually to perform the computations.

Taking eqn (7) into account and neglecting the unimportant difference between the partial derivatives at the points  $a_i$  and  $a_{i0}$  the parameter deviation can be written as follows:

$$da_i = \frac{n \overline{\left( \frac{\partial q}{\partial a_i} \right)_{a_{i0}}}}{\overline{\left( \frac{\partial q}{\partial a_i} \right)_{a_{i0}}^2}} \quad (8)$$

After introduction of the model weighting function the following alternative expression for the parameter deviation is obtained.

$$da_i = \frac{\int_{-\infty}^{\infty} du \left( \frac{\partial Q(u)}{\partial a_i} \right)_{a_{i0}} \overline{n(t) x(t-u)}}{\int_{-\infty}^{\infty} du \int_{-\infty}^{\infty} dv \left( \frac{\partial Q(u)}{\partial a_i} \right)_{a_{i0}} \left( \frac{\partial Q(v)}{\partial a_i} \right)_{a_{i0}} \overline{x(t-u) x(t-v)}} \quad (9)$$

The expression eqn (9) relates a certain parameter deviation to properties of the model weighting function, the input signal and the process noise. The time interval  $T_m$  of the test is inherent in the averaging procedures indicated by the bars. The expression is valid for a parameter isolated by the requirements of orthogonality.

### Estimation of the Parameter Deviation R.M.S. Value

A stationary process noise source with a zero mean output makes the average deviation  $da_i$  zero, because positive and negative outcomes of the same magnitude are equally likely. Considering the mean square value, it is convenient to treat the numerator ( $r$ ) and the denominator ( $d$ ) of eqn (9) separately, the former being a stochastic quantity and the latter a deterministic one.

The auxiliary function  $p(t)$  allows us to write the numerator  $r$  as follows:

$$p(t) = \int_{-\infty}^{\infty} du \left( \frac{\partial Q(u)}{\partial a_i} \right)_{a_{i0}} x(t-u) \quad (10)$$

$$r = \overline{p(t) n(t)} \quad (11)$$

In the next step, the numerator will be squared and the bars exchanged by the corresponding integral signs. Simultaneously a time displacement  $v$  of the noise signal will be introduced as a tool to shift any interval of the endless stream from the noise

source into the range of measurement  $T_m$ . An index  $p$  is attached to the symbol  $r$  to indicate that the result is valid for a certain function  $p(t)$ .

$$r_p^2(v) = \frac{1}{T_m^2} \int_{-\infty}^{\infty} dt_1 \int_{-\infty}^{\infty} dt_2 p(t_1) p(t_2) n(v+t_1) n(v+t_2) \quad (12)$$

The average, with respect to  $v$ , of the squared numerator is obtained if the noise signals under the integral signs are exchanged by the noise auto-correlation function  $\varphi_{nn}(t_2 - t_1)$ . For convenience the difference  $t_2 - t_1$  is also replaced by  $t$  giving the following result.

$$\overline{r_p^2} = \frac{1}{T_m^2} \int_{-\infty}^{\infty} dt \left[ \int_{-\infty}^{\infty} dt_1 p(t_1) p(t_1+t) \right] \varphi_{nn}(t) \quad (13)$$

The function within the brackets is also an auto-correlation function and may be denoted by  $\varphi_{pp}(t)$ . This correlation should be computed on the basis of the recorded data about  $x(t)$ . In the case, however, where the input signal is generated by a stochastic source the correlation function might be computed directly from the source characteristics. It is therefore sufficient to know the input signal in statistical terms only. The mean square value of the numerator  $r$  can therefore be written as follows, both for deterministic and stochastic input signals.

$$\overline{r^2} = \frac{1}{T_m} \int_{-\infty}^{\infty} dt \varphi_{pp}(t) \varphi_{nn}(t) \quad (14)$$

Often the signal power density spectrum,  $\Phi_{xx}(\omega)$ , can preferably be used in the calculation of the integral eqn (14). According to eqn (10) we have:

$$\Phi_{pp}(\omega) = \Phi_{xx}(\omega) \left| \left( \frac{\partial Q(u)}{\partial a_i} \right)_{a_{i0}} \right|^2 \quad (15)$$

Parseval's formula applied to eqn (14) gives then the following alternative expression for the numerator mean square value.

$$\overline{r^2} = \frac{2\pi}{T_m} \int_{-\infty}^{\infty} d\omega \Phi_{xx}(\omega) \Phi_{nn}(\omega) \left| \left( \frac{\partial Q(j\omega)}{\partial a_i} \right)_{a_{i0}} \right|^2 \quad (16)$$

The denominator  $d$  of the eqn (9) is independent of the noise and has a definite value when the input signal  $x(t)$  is given. This value is equal to the mean square value of  $p(t)$ , as can be seen by comparing the denominator and the definition of  $p(t)$ , eqn (10).

$$d = \overline{p^2(t)} \quad (17)$$

In the case of stochastic input signals, use can be made of the eqn (15) giving this alternative expression for the denominator.

$$d = \varphi_{pp}(0) = \int_{-\infty}^{\infty} d\omega \Phi_{xx}(\omega) \left| \left( \frac{\partial Q(j\omega)}{\partial a_i} \right)_{a_{i0}} \right|^2 \quad (18)$$

Finally we are able to express the parameter deviation R.M.S. value, or the parameter uncertainty, in one single relation by the introduction of eqns (16) and (18) into eqn (9).

$$\overline{da_i^2} = \frac{2\pi \int_{-\infty}^{\infty} d\omega \Phi_{xx}(\omega) \Phi_{nn}(\omega) \left| \left( \frac{\partial Q(j\omega)}{\partial a_i} \right)_{a_{i0}} \right|^2}{\left[ \int_{-\infty}^{\infty} d\omega \Phi_{xx}(\omega) \left| \left( \frac{\partial Q(j\omega)}{\partial a_i} \right)_{a_{i0}} \right|^2 \right]^2} \quad (19)$$

The uncertainty relation is here presented in the frequency domain. It can readily be written also in the time domain by use of eqns (14) and (17). The relation will then, however, be somewhat more complicated due to the two convolution integrals concealed in  $\varphi_{pp}(t)$ . Numerical and graphical calculations seem, in general, to be easier to perform in the frequency domain.

### First Example: A Frequency Response Test

In the first example, a sinusoidal input signal,  $a \sin \nu t$ , is used to determine the three parameters  $K$ ,  $T$  and  $L$  of a simple process, described by the following transfer function:

$$G(j\omega) = \frac{K e^{-j\omega L}}{1 + j\omega T} \quad (20)$$

For a numerical illustration these specific values of the parameters,  $K = 1$ ,  $L = T$  are chosen.

The noise power density spectrum is assumed to be constant over the frequency range of interest.

$$\Phi_{nn}(\omega) = N_0 \quad \text{unit}^2/\text{rad/sec} \quad (21)$$

$$N_0 = \frac{a^2 T}{400} \quad (22)$$

The noise power of eqn (22) corresponds to a ratio 1/10 between the R.M.S. value of noise components within the frequency band  $-1/T$  to  $+1/T$  and the R.M.S. value of the sinusoidal input.

From eqn (20) the following partial derivatives are obtained.

$$\begin{aligned} G_K(j\omega) &= \frac{e^{-j\omega L}}{1 + j\omega T} \\ G_T(j\omega) &= \frac{-j\omega K e^{-j\omega L}}{(1 + j\omega T)^2} \\ G_L(j\omega) &= \frac{-j\omega K e^{-j\omega L}}{1 + j\omega T} \end{aligned} \quad (23)$$

The uncertainty relation, eqn (19), is now simplified by the facts that the noise spectrum is a constant and the area of the input signal spectrum is concentrated at the frequencies  $\pm \nu$ . The remaining integral to be solved, the value of which is given by the definition of the spectrum, is as follows:

$$\int_{-\infty}^{\infty} d\omega \Phi_{xx}(\omega) = \frac{a^2}{2} \quad (24)$$

The substitution of eqns (21), (23) and (24) into the relation eqn (19) yields the following result:

$$\frac{\overline{dK^2}}{K^2} = \frac{4\pi}{T_m} \frac{N_0}{K^2 a^2} (1 + \nu^2 T^2) \quad (25a)$$

$$\frac{\overline{dT^2}}{T^2} = \frac{4\pi}{T_m} \frac{N_0}{K^2 a^2} \left( \nu T + \frac{1}{\nu T} \right)^2 \quad (25b)$$

$$\frac{\overline{dL^2}}{L^2} = \frac{4\pi}{T_m} \frac{N_0}{K^2 a^2} \left( \frac{T}{L} \right)^2 \frac{1 + \nu^2 T^2}{\nu^2 T^2} \quad (25c)$$

A low frequency is preferable in the determination of the gain  $K$ , a high frequency better for the lag  $L$ , and the specific frequency  $\nu = 1/T$  best with respect to the time constant  $T$ . Choosing the latter frequency and introducing the numerical data given above one gets

$$\begin{aligned}\frac{(\overline{dK^2})^{\frac{1}{2}}}{K} &= 0.25 \left( \frac{T}{T_m} \right)^{\frac{1}{2}} \\ \frac{(\overline{dT^2})^{\frac{1}{2}}}{T} &= 0.35 \left( \frac{T}{T_m} \right)^{\frac{1}{2}} \\ \frac{(\overline{dL^2})^{\frac{1}{2}}}{L} &= 0.25 \left( \frac{T}{T_m} \right)^{\frac{1}{2}}\end{aligned}\quad (26)$$

To determine the parameters  $K$ ,  $T$ , and  $L$  within a relative accuracy of 5 per cent, the measuring time  $T_m$  should be at least  $25 T$ ,  $50 T$  and  $25 T$  respectively.

The result, eqn (26), is independent of the evaluation procedure, provided an optimal one is employed. In one conventional method of frequency response testing the output is multiplied by a sine and a cosine function and, after filtering, the 'in-phase' and 'quadrature' components are recorded. This method is optimal in the sense assumed here. A direct calculation of the parameter errors, when the parameters are derived in the best possible way from the in-phase and quadrature components, will give exactly the result of eqns (25) and (26). The calculation is too lengthy to be included in this paper. Reference is made instead to the well-known fact that the conventional frequency response method is a kind of Fourier transformation, which can be developed from the least mean square error criterion being our starting point.

## Second Example: A Random Input Signal

The same process, eqn (20), is studied in this example but this time a random binary input signal is employed. The input is constant  $+a$  or  $-a$  during successive and equal intervals  $T_0$ , the probability of each sign being  $1/2$ .

The evaluation of the parameters  $K$ ,  $T$  and  $L$  can be done according to the scheme indicated in the introduction; a model adjustment by means of the recorded input and output signals.

The process transfer function can also be obtained via the Wiener-Hopf integral equation, if appropriate correlation functions are computed by use of the recorded data. The solution of the integral equation fulfils by definition the same least mean square error criterion used in this paper. The uncertainty relation, eqn (19), should therefore be applicable. The deduction of this equation was, in fact, in the beginning of this work, based upon the Wiener-Hopf equation, before a direct approach was found preferable.

The auto-correlation function  $\varphi_{xx}(\tau)$  of the binary signal can be found in several textbooks and the power density spectrum is readily obtained by Fourier transformation.

$$\begin{aligned}\varphi_{xx}(\tau) &= a^2 \left( 1 - \frac{|\tau|}{T_0} \right) \\ \Phi_{xx}(\omega) &= \frac{a^2}{\pi T_0} \frac{1 - \cos \omega T_0}{\omega^2}\end{aligned}\quad (27)$$

In the present example, the spectrum of the input is distributed and the partial derivative functions, eqn (23), will re-

main under the integral sign when eqn (23) is substituted for eqn (19). This implies, however, only an unimportant complication, as the integral can be solved in a straightforward way by contour integration (if not found in a list of definite integrals).

The uncertainty relations will take the following form after the integrations have been carried out.

$$\begin{aligned}\frac{\overline{dK^2}}{K^2} &= \frac{2\pi}{T_m} \frac{N_0}{K^2 a^2} \frac{T_0/T}{T_0/T - 1 + e^{-T_0/T}} \\ \frac{\overline{dT^2}}{T^2} &= \frac{4\pi}{T_m} \frac{N_0}{K^2 a^2} \frac{T_0/T}{1 - (1 + T_0/T) e^{-T_0/T}} \\ \frac{\overline{dL^2}}{L^2} &= \frac{2\pi}{T_m} \frac{N_0}{K^2 a^2} \left( \frac{T}{L} \right)^2 \frac{T_0/T}{1 - e^{-T_0/T}}\end{aligned}\quad (28)$$

A long time interval  $T_0$  favours the determination of the gain  $K$ , the shortest possible interval will suit the parameter  $L$  and a certain interval  $T_0 = 1.8 T$  is the best choice with respect to the time constant  $T$ . This value of the interval will be accepted for the following numerical example.

As in the first example a ratio of  $1/10$  will be assumed between the R.M.S. value of the noise components in the band  $-1/T$  to  $+1/T$  and the input signal R.M.S. value.

$$N_0 = \frac{a^2 T}{200}\quad (29)$$

The introduction of eqn (29) and the values  $T_0/T = 1.8$ ,  $L = T$  and  $K = 1$  into eqn (28) will give this result.

$$\begin{aligned}\frac{(\overline{dK^2})^{\frac{1}{2}}}{K} &= 0.24 \left( \frac{T}{T_m} \right)^{\frac{1}{2}} \\ \frac{(\overline{dT^2})^{\frac{1}{2}}}{T} &= 0.46 \left( \frac{T}{T_m} \right)^{\frac{1}{2}} \\ \frac{(\overline{dL^2})^{\frac{1}{2}}}{L} &= 0.26 \left( \frac{T}{T_m} \right)^{\frac{1}{2}}\end{aligned}\quad (30)$$

The measuring time  $T_m$  to get a relative accuracy of 5 per cent should here be at least  $23.5 T$ ,  $85 T$  and  $27 T$  for  $K$ ,  $T$  and  $L$  respectively.

The parameter uncertainties of the stochastic case, eqn (30), can directly be compared with the result of the preceding example of a deterministic input signal. The general conditions are the same in both cases. No discrepancy can be observed with respect to the result.

## Conclusion

The derivation of this paper resulted in a relationship between the uncertainty of an arbitrary model parameter and four important factors of the identification problem. These factors are: the input signal and the process noise, represented by their power density spectra, the time period of measuring and the influence of the parameter being considered upon the model weighting or transfer function. The parameter influence is measured by the partial derivative of the latter function with respect to the parameter of interest.

It is possible to express the uncertainty relation in the time or in the frequency domain. The frequency domain presentation, eqn (19), describes the main features of the relation in an easier way and is more suitable for approximate and graphical computations.

It will be of interest to sum up the qualities of the factors mentioned above, that will tend to reduce the uncertainty of the model.

- (a) The input signal should be large, and located as much as possible within a frequency range defined by the parameter influence function.
- (b) The process noise should be small, and located as much as possible outside the frequency range given in (a).
- (c) A long time of measurement is favourable and the uncertainty is inversely proportional to the square root of that time.

Although the present result in the first place is useful in the planning of transfer function measurements, it illustrates also a dilemma of adaptive control. In such a control system a process parameter is continuously evaluated from process input and output data to initiate an adjustment of the controller, compensating for parameter variations. The time of measurement is equivalent to a time delay in the loop of adaptation; a short time delay means therefore a fast, but incorrect, compensation. A long time of measurement allows the compensation to be correct in size, but too late in time. Thus the uncertainty relation indicates that some unfavourable parameter variations cannot even theoretically be compensated for by adaptive control.

## Nomenclature

$x(t)$	The process (and model) input signal
$y(t) + n(t)$	The process output signal
$y(t)$	The deterministic part of the process output signal
$n(t)$	The stochastic part of the process output signal
$q(t)$	The model output signal

$$e(t) = y(t) + n(t) - q(t)$$

$$T_m$$

$$\overline{e^2(t)}$$

$$t, \tau, u, v$$

$$Q(a_i, \tau)$$

$$Q(j\omega)$$

$$a_i$$

$$N$$

$$\varphi_{nn}(\tau) = \lim_{T \rightarrow \infty} \frac{1}{2T} \int_{-T}^T n(t) n(t + \tau) dt$$

$$\Phi_{nn}(\omega) = \frac{1}{2\pi} \int_{-\infty}^{\infty} d\tau e^{-j\omega\tau} \varphi_{nn}(\tau)$$

$$X(j\omega) = \int_{-\infty}^{\infty} x(t) e^{-j\omega t} dt$$

$$x(t) = \frac{1}{2\pi} \int_{-\infty}^{\infty} X(j\omega) e^{j\omega t} d\omega$$

$$\begin{aligned} \int_{-\infty}^{\infty} u(t) v(t) dt \\ = \frac{1}{2\pi} \int_{-\infty}^{\infty} U(j\omega) V(-j\omega) d\omega \end{aligned}$$

The difference between the process and the model output signals

The period of time during which the process input and output signals have been recorded

The bar indicates the average with respect to time over an appropriate time interval, usually  $T_m$

Time variables

The model weighting function

The model transfer function or the Fourier transform of  $Q(t)$

The parameter number  $i$  of the model weighting or transfer function

The total number of parameters  $a_i$

The auto-correlation function of the process noise  $n(t)$

The power density spectrum of  $n(t)$

Definition of the Fourier transform

The Parseval theorem

## DISCUSSION

P. EYKHOFF, *Technological University, Electrical Engineering Department, Delft, Netherlands*

The contribution by Dr. Qvarnström deserves to be studied by all who are concerned with process-parameter estimation.

The time intervals needed for information processing that result from his studies clearly indicate the basic limitations inherent in 'tracking' of unknown parameters.

In the paper, use is made of orthogonality, which leads me to the following questions:

(1) In eqn (7) of the paper, a condition of orthogonality is stated. For stochastic types of input signals this condition is difficult to fulfil; the construction of a set of filters orthogonal with respect to a single, time limited element of a stochastic process seems to be highly impractical. So the best we can hope for is

$$E \left[ \left( \frac{\partial q}{\partial a_i} \right)_{a_i} \left( \frac{\partial q}{\partial a_j} \right)_{a_{j0}} \right] = 0 \quad j \neq i \quad (1)$$

where  $E[\ ]$  indicates the mathematical expectation. Do you have an estimate of how much the results that are derived in the paper degrade as a result of this change of orthogonality conditions?

(2) Just below eqn (7) of the paper is stated 'The orthogonal parameter can be constructed as linear combinations of the original parameters plus extra, auxiliary parameters'. What are the theoretical and practical limitations of this procedure with respect to the parameter interval over which it holds?

(3) In spite of the conditions of orthogonality introduced in the derivation of the paper's eqn (19) one notices in the example given that orthogonality of the parameters is not fulfilled. What are the implications of this with respect to the answer obtained?

(4) On the third page of the paper the idea of an optimal evaluation procedure is coined. In engineering terms, what does constitute such an optimal procedure?

F. B. TUTEUR, *Department of Engineering and Applied Science, Yale University, New Haven, Connecticut, U.S.A.*

The author is to be congratulated for his simple and straightforward application of modern decision theory to the problem of system identification. The results appear to be quite universal, and they place a lower bound on the errors that can be obtained from measurements made during a finite time. Thus, they would appear to settle once and for all the question of whether it is possible to find an identification

scheme that is very much superior to others that have been considered in the past. It is interesting to note that the R.M.S. error varies inversely with the square root of the estimation time. This is, of course, in accordance with typical results in estimation theory and serves as a qualitative check.

Since the author has confined himself to the estimation of a known set of parameters, the paper does not consider the problem of identification of a system whose exact structure is not known. I would, however, be interested in the author's opinion about the possibility of applying his method in a situation where the structure is not known, or only partially known, i.e. where the system has more or different parameters than the model.

E. BLANDHOL, *Eidanger Salpeterfabriker, Herøya, Porsgrunn, Norway*

Two very important points in the practical use of model methods are the choice of model structure and the convergence of the experimental procedures.

The author makes the following two assumptions: (a) that the model and process are identical for a certain set of parameters  $a_{i0}$ , and (b) that the model parameters are orthogonal. Assumption (a) can only be satisfied if the system structure is exactly known and if it is possible to duplicate it in the model. In a mathematical model of a 'black box' system the assumption will probably be invalid. Assumption (b) of orthogonal model parameters may be satisfied in a mathematical model, but usually not in a model duplicating a given system structure. I would like to ask the author how important the two assumptions are for the validity of his results.

The convergence of the model adjustment depends on the form of the 'performance surface', and not only in the vicinity of optimum, where the author's formulae are valid, but in a broader range. In the paper by J. G. Balchen and myself we observed, by our experiments, that for certain model structures and adjustment strategies the method may not converge at all. The existence of convergence must therefore be known before the author's results can be applied in a practical case to judge the results.

In his derivations the author introduces the statistical properties of the input and the noise, thereby arriving at measurement times that seem unnecessarily long. It is quite possible to use very short test signals of a simple form, such as a step or ramp, and still obtain useful results. In our paper we have studied both first-order and second-order systems with output noise, and the true values, with a measuring time of only  $8T$ , where  $T$  is the largest time constant in the system. This seems to be somewhat better than predicted by the author.

J. E. RIJNSDORP, *Koninklijke/Shell Laboratorium, Amsterdam, Netherlands*

In your first example you consider a frequency response test. Should not the time for completion of the transient response to the sinusoidal test signal be added to the measuring time  $T_m$ ?

B. QVARNSTRÖM, *in reply*

In his first question Dr. Eykhoff emphasizes a problem arising from the fact that the orthogonality condition is dependent on the input

signal. I am not at present prepared to give a precise and general statement regarding the consequences of that dependence, but would like to say the following:

(1) The result is not very sensitive to the effect of coupling between different parameters, as demonstrated in my oral presentation. Therefore, a moderate lack of orthogonality should not upset the result.

(2) Although it might be possible to construct a case in which the input signal in a critical manner will affect the condition of orthogonality, I have not found such a case when a stochastic signal has been employed.

The model adjustment procedure, being the basis of my work and of others, can certainly be regarded as a very good one, because all information available is utilized in forming the model. Accepting the mean square error criterion I consider, by definition, this procedure to be an optimal one. There are other procedures giving exactly the same parameter uncertainty; these should, therefore, also be optimal. One example is the frequency response test, when use is made of the Fourier or correlation method.

Dr. Tuteur and Mr. Blandhol are concerned with the problem of processes of unknown structure. Although not clearly stated in my paper, it is fairly obvious that the result can be used in the case of such processes, provided the following interpretation of the uncertainty is made. The formula eqn (19) gives the minimum uncertainty of a specific model parameter that can be obtained by the best possible use of the measured data. It is not necessary that a corresponding process parameter exists. The process or system may therefore have more or different parameters than the model.

If the parameter under consideration is coupled to other parameters of the system, this effect will on the average increase the uncertainty above the minimum value. Here we have a choice, orthogonal parameters can be used, which is theoretically possible, or the coupling can be taken into account by a more complex uncertainty formula.

Mr. Blandhol brings up the important question about the convergence of the model adjustment procedure. That problem is outside my work. When a suitable procedure is found, my formula will indicate how much the parameter values obtained are affected by the process noise. In one case of interacting parameters I found that the procedure did not converge, but at the same time the appropriate uncertainty formula (not given in the present paper) indicated infinite parameter uncertainties.

It is not quite correct that statistical properties of the input signal are introduced. My result is valid also for deterministic input functions and reveals that under favourable circumstances a short measuring time is possible. I have checked my uncertainty estimation with the result of the paper by Messrs. Blandhol and Balchen and found a reasonable agreement.

Mr. Rijnsdorp is quite right. There might, however, be several different situations. The sinusoidal signal might be started and turned off within the measuring period, which should then include the transients also. In the case of a continuous sinusoidal signal the transients will be lost; it is then necessary that the measuring time is long compared with the time constants of the process.

# Axiomatic Foundation of the Theory of Control Systems

E. ROXIN

## Summary

The purpose of this paper is to show how an axiomatic development of the theory of control systems can systematize important concepts. Starting with the idea of attainable set (the set of points which can be reached from a given one in a certain time) a set of 6 axioms is given to describe the behaviour of a general control system. Barbashin already gave such an axiomatic approach, and Zubov also used an axiomatic theory for proving stability theorems. But besides the strong form of stability and invariance of a set (given by Zubov) a weak form of these properties can be defined for general control systems. In general, almost every property of the classical results can be translated in a strong and a weak form for control systems. Accordingly, the powerful second method of Liapunov can also be applied in a strong and a weak form, in order to prove strong and weak properties, respectively.

In this paper is given the set of basic axioms (more or less equivalent to Barbashin's); some basic results are discussed; the strong and weak form of invariance and stability properties and the corresponding strong and weak Liapunov functions are defined, giving the statement of the fundamental stability theorems (the detailed proofs are given elsewhere).

## Sommaire

Dans ce travail on montre comment un développement axiomatique permet d'exposer systématiquement d'importantes notions de la théorie des systèmes de commande. Avec la notion d'ensemble accessible (l'ensemble des points qu'on peut atteindre d'un point donné en un certain temps), on peut caractériser les systèmes de contrôle par 6 axiomes. Barbashin a donné un système d'axiomes analogues et Zubov a utilisé un système axiomatique pour démontrer des théorèmes concernant la stabilité. Mais parallèlement à la forme forte de la stabilité et de l'invariance d'un ensemble, utilisée par Zubov, on peut définir une forme faible. En général, la plupart des propriétés classiques des systèmes dynamiques peut être traduite d'une manière forte ou faible pour les systèmes de commande. La seconde méthode de Liapunov est aussi applicable aux formes forte et faible.

Dans ce travail on donne un ensemble d'axiomes (semblables à celles données par Barbashin), les principales conséquences sont discutées, on définit les formes faible et forte de notions des stabilité et d'ensemble invariant, de plus les fonctions de Liapunov on sens fort et faible et l'on énonce les principaux théorèmes (les démonstrations seront données ailleurs).

Finalement on discute les relations avec des résultats classiques (conditions de différentiabilité) et les possibles développements ultérieurs.

## Zusammenfassung

Zweck dieses Beitrages ist es, aufzuzeigen, wie eine axiomatische Entwicklung der Regelungstheorie wichtige Begriffe systematisieren kann. Ausgehend vom Begriff der „erreichbaren Menge“ (einer Menge von Punkten, die in bestimmter Zeit von einem gegebenen Punkt ausgehend erreicht werden können) werden sechs Axiome angegeben, um das Verhalten eines allgemeinen Regelungssystems zu beschreiben. Barbashin hat bereits einen derartigen axiomatischen Weg angegeben und auch Zubov verwendete eine axiomatische Theorie, um Stabilitätstheoreme zu beweisen. Aber neben der strengen Form der Stabilität und der Invarianz einer Menge (bei Zubov) kann eine schwache Form dieser Eigenschaften für allgemeine Regelungssysteme definiert

werden. Im allgemeinen läßt sich fast jede Eigenschaft der klassischen Ergebnisse für Regelungssysteme in eine strenge und in eine schwache Form übersetzen. Entsprechend kann auch die wirkungsvolle zweite Methode von *Ljapunow* in einer strengen und in einer schwachen Form angewendet werden, um entsprechend „strenge“ oder „schwache“ Eigenschaften zu beweisen.

Der Beitrag enthält die grundlegenden Axiome (sie sind mehr oder weniger denen von Barbashin gleichwertig), die Diskussion einiger grundsätzlicher Ergebnisse, die strengen und schwachen Formen der Invarianz und der Stabilitätseigenschaften sowie eine Definition der zugehörigen strengen und schwachen *Ljapunow*-Funktionen, welche zu den grundsätzlichen Stabilitätstheoremen führen. Die ins einzelne gehenden Beweise sind anderenorts angegeben.

Schließlich werden der Zusammenhang mit den klassischen Ergebnissen (Bedingungen der Differenzierbarkeit) sowie einige weitere Entwicklungen kurz erwähnt.

## Introduction

From its beginning, the theory of control systems developed as a branch of the theory of differential equations. More general approaches, including differential difference and integral equations, are also important, but the most developed theory is related to the classical differential equations. The problems of control theory appear, this way, as special applications of this theory; but in a certain sense the problem of control theory can be regarded as more general than the one corresponding to differential equations, as will be seen. Now, the theory of differential equations has been successfully developed and generalized to the theory of dynamical system<sup>2, 3</sup>. Instead of starting with the equations, one assumes a set of axioms which characterize the most basic properties of the solution curves, obtaining all the desired results directly from those axioms. The advantage is that, in this way, the knowledge and even the existence of the differential equations governing the motion, is not assumed, so that, for example, the differentiability conditions for the solutions can be relaxed. There is a temptation, therefore, to develop the theory of control systems along the same lines.

A control system, given normally by a differential equation of the type

$$\dot{x} = f(x, t, u) \quad (1)$$

where  $x$  is an  $n$  vector (the state variable),  $t$  is a real variable (the time), and  $u$  is an  $m$  vector (the control or steering function), can be regarded as a more general system than a simple differential equation. Indeed, such an equation assigns to each point  $x$  and time  $t$  a definite direction vector; meanwhile the control system (1) assigns to the couple  $x, t$  a whole set of possible values of  $\dot{x}$ , according to the choice of the control parameter  $u$ . Systems, where at each point a whole cone or a more general set of possible directions are given, were considered



a long time ago by Zaremba<sup>8</sup> and Marchaud<sup>5</sup>. These authors obtained the existence theorems and basic results concerning such systems.

The next step of giving a set of axioms governing the motions defined by control systems, in order to develop the theory independently from equation (1), was undertaken by Barbashin<sup>1</sup>. Among later applications of this same idea, Zubov<sup>9</sup> must be mentioned; he gave stability theorems using Liapunov-type functions for systems defined axiomatically (using a different set of axioms to Barbashin); but apparently these works were more influenced by the classical results about dynamical systems, trying to include the case of non-unique solutions through each point rather than by the problems arising in control theory. In this paper it is attempted to show that with this axiomatic approach one can obtain results of important significance in control theory.

### Notation

Consider a complete locally compact metric space  $X = \{x\}$  of the state variable  $x$ . The real variable  $t \in R = (-\infty, +\infty)$  is called time. In order to simplify the ideas only autonomous systems are considered, but most results can easily be generalized for time-dependent systems.

Point sets in  $X$  are denoted by capital letters  $A, B, \dots$ ; the elements of  $X$  and also real numbers or functions, by small letters  $x, y, \dots$ . The distance defining the metric in  $X$  is  $\rho(x, y)$ . Sometimes, in order to avoid infinite distance between sets, this metric can be replaced by  $\rho(x, y) / (1 + \rho(x, y))$ , but in this paper no difficulty will arise in this aspect. The distance between a point  $x$  and a set  $A$  is defined by  $\rho(x, A) = \rho(A, x) = \inf \{\rho(x, a); a \in A\}$ . 'Separation' of the set  $A$  from the set  $B$  is called the number  $\rho^*(A, B) = \sup \{\rho(a, B); a \in A\}$ . It should be noted that this separation is not symmetric in  $A, B$ ; in general, that is  $\rho^*(A, B) \neq \rho^*(B, A)$ . The 'distance' between sets  $A$  and  $B$  is then  $\rho(A, B) = \max \{\rho^*(A, B); \rho^*(B, A)\}$ . In the space of the compact subsets of  $X$ , this distance function defines a true metric<sup>4, 5</sup>. If one admits closed but not compact subsets, this distance can become infinite (and this can be avoided by the use of the above-mentioned change of the distance function, which does not change the topology). If one also wishes to take into account non-closed sets, it defines only a pseudo-metric. As compact subsets are dealt with, further details will not be given.

Finally,  $S_r(A) = \{x; \rho(x, A) \leq r\}$  will denote the  $r$ -neighbouring set of  $A$ ; note that this is a subset of  $X$  and not a neighbourhood of  $A$  considered as an element of a space of subsets.

### The Attainability Function

#### Basic Axioms

The state-point  $x \in \{X\}$  will change with time  $t$ , according to conditions inherent to the system and to some specific control action. Before choosing any particular control function, the only thing which can be stated is the whole set of points of  $X$  which are attainable, in the time  $t$ , from a given  $x$ . So the system is characterized, giving a function  $F(x, t)$ , which is a subset of  $X$ , with the meaning  $y \in F(x, t)$  if, and only if, starting at  $x$  for  $t = 0$ , one can reach the point  $y$  at the time  $t$  by a suitable choice of the control function.

Disregarding formally that meaning of the attainability function, it is assumed that the following axioms are fulfilled:

(I)  $F(x, t)$  is defined for every  $x \in X, t \geq 0$ , and is a closed non-void set.

(II) Initial condition:  $F(x, 0) = \{x\}$  for every  $x$ .

(III) Semigroup property: if  $0 \leq t_1 \leq t_2$ , then

$$F(x_0, t_2) = \bigcup_{x_1 \in F(x_0, t_1)} F(x_1, t_2 - t_1)$$

(IV) Given  $x_1, t_1 \geq 0$ , there exists (at least one)  $x_0$  such that  $x_1 \in F(x_0, t_1)$ .

(V) Continuity with respect to  $t$ :  $F(x, t)$  is continuous in  $t$ . This means that, given  $x_0, t_0 \geq 0$  and  $\varepsilon > 0$ , there is a  $\delta > 0$  such that for any  $t \geq 0, |t - t_0| < \delta$ , the inequality  $\rho[F(x_0, t_0), F(x_0, t)] < \varepsilon$  is satisfied.

(VI) Upper semicontinuity with respect to  $x$ : Given  $x_0, t_0 \geq 0$  and  $\varepsilon > 0$ , there is a  $\delta > 0$  such that for any  $x \in X$ , with  $\rho(x, x_0) < \delta$ , the inequality  $\rho^*[F(x, t_0), F(x_0, t_0)] < \varepsilon$  is satisfied.

By examples it can be shown that these axioms are independent of each other. It is interesting to note that axiom (III) is equivalent to the following two, which is formulated only in words:

(III a) If  $x_1$  is attainable from  $x_0$  in the time  $t_1$ , and  $x_2$  is attainable from  $x_1$  in the time  $t_2$ , then  $x_2$  is attainable from  $x_0$  in the time  $t_1 + t_2$ .

(III b) If  $x_2$  is attainable from  $x_0$  in the time  $t_2$ , and  $0 \leq t_1 \leq t_2$ , then there is some point  $x_1$  attainable from  $x_0$  in time  $t_1$ , such that from it  $x_2$  is attainable in time  $t_2 - t_1$ .

Barbashin<sup>1</sup> gives an equivalent set of axioms, and assumes the function  $F(x, t)$  defined for all real values of  $t$ . It seems, perhaps, more natural to the essence of control theory to start assuming only positive values of  $t$  (evolution to the future, disregarding the past). Zubov<sup>9</sup> also considers this kind of systems for positive times, even without stating explicitly these axioms.

In a straightforward way the set function

$$F(A, t) = \bigcup_{x \in A} F(x, t)$$

can be defined; for this function properties similar to axioms (I) to (VI) hold. If one only considers compact sets  $A$ , the similarity is complete, but if allowing all non-void subsets of  $X$ , then it must be remembered that they constitute only a pseudo-metric space, which is not Hausdorff<sup>6</sup>.

The systems satisfying these axioms include, for example, the differential systems which fail to satisfy the condition of unique solution (i.e. the Lipschitz condition). If, instead of the upper semicontinuity of axiom (VI), regular continuity of  $F(x, t)$  with respect to  $x$  is required, one excludes these systems, but still includes the technical important control problems.

#### Consequences of the Axioms

**Trajectories**—The most remarkable consequences of the given axioms follow.

The set  $F(x, t)$  is compact. If  $A$  is compact, then so is  $F(A, t)$ . This is a consequence of the continuity axiom (V). It rules out the possibility of solutions with a finite escape time.

One can extend the domain of the attainability function to negative values of  $t$ , defining:  $y \in F(x, -t)$  if, and only if,  $x \in F(y, t)$  ( $t > 0$ ). All but the continuity conditions are then

also satisfied for negative values of  $t$ . On the other hand, it is possible to have 'finite escape times' for negative values of  $t$ , so that  $F(x, t)$  is no longer compact. To avoid this, one has to assume an additional hypothesis, but it is not necessary for the further development of the theory.

A trajectory is defined as a curve  $x(t)$  in some  $t$  interval, such that if  $t_1 < t_2$  belong to that interval,  $x(t_2) \in F[x(t_1), t_2 - t_1]$ . A fundamental theorem states that if  $y \in F(x, t)$ , then there exists some trajectory going from  $x$  to  $y$  (supposed  $t > 0$ ). The notation  $x = \phi(x_0, t)$  is used here for a trajectory. Barbashin has proved the following strong theorem: if  $x_n \rightarrow x_0$  and  $\phi_n(x_n, t)$  is a sequence of trajectories through  $x_n$ , then there exists a subsequence tending (pointwise) to some limit trajectory passing through  $x_0$ :  $\phi_n(x_n, t) \rightarrow \phi_0(x_0, t)$ .

### Strong and Weak Properties of Control Systems

Now one is in a position to extend to control systems all the properties normally found in dynamical systems. There are two very important ways to generalize any property. Suppose that in the theory of ordinary dynamical systems a set is called ' $N$ ' if the trajectories starting at points of that set have the property ' $p$ '. In the ordinary dynamical systems there is one well-defined trajectory through each point, so that this definition is precise. In control systems, instead, there is a whole set of trajectories through each point, which makes the direct application of the same definition ambiguous. It will then be said that the set  $A$  is 'strongly  $N$ ' if all the trajectories starting at points of  $A$  have the property ' $p$ '. On the other side it will be said that the set is 'weakly  $N$ ' if, starting at each point of  $A$ , there is *some* trajectory which has the property ' $p$ '. To take an example: the set  $A$  is called strongly positively invariant if every trajectory starting at a point of  $A$  remains in  $A$  for all positive  $t$ . It is weakly invariant if at each point of  $A$  there starts (at least) one trajectory remaining in  $A$  for all  $t > 0$ . The meaning is obvious: if  $A$  is weakly invariant, one is able to remain in  $A$  by a convenient choice of the trajectory; if, instead,  $A$  is strongly invariant, one cannot leave  $A$  for any choice of the trajectory. In a similar way all stability definitions of dynamical systems give rise to strong and weak definitions in control systems.

In order to study stability properties Liapunov's functions for control systems are defined, considering upper and lower total generalized derivatives for those functions which correspond to the strong and weak properties to be proved.

### Definitions

All the following definitions refer to a certain general control system, defined by the attainability function  $F(x, t)$ .

The set  $A$  is a strongly *positively* invariant set if  $x \in A \Rightarrow F(x, t) \subset A$  for all  $t \geq 0$ .

The set  $A$  is a strongly *negatively* invariant set if  $x \in A \Rightarrow F(x, t) \subset A$  for all  $t \leq 0$ .

The set  $A$  is a 'strongly invariant set' if  $x \in A \Rightarrow F(x, t) \subset A$  for all  $t$ .

The set  $A$  is 'weakly positively invariant' if for any  $x \in A$  there is a trajectory  $\phi(x, t)$  contained in  $A$  for all  $t \geq 0$ . It can be seen<sup>1</sup> that this requirement is equivalent to the following:  $A$  is positively weakly invariant if  $x \in A \Rightarrow F(x, t) \cap A \neq \emptyset$  for  $t \geq 0$ . Similar definitions hold for negatively weakly invariant and weakly invariant sets.

The set  $A$  will be called 'strongly stable' if for any  $\varepsilon > 0$  there is a  $\delta > 0$  such that  $x \in S_\delta(A) \Rightarrow F(x, R^+) \subset S_\varepsilon(A)$ , where  $R^+ = (0, \infty)$ . *Remark:* If  $A$  is strongly stable then closure  $A$  is strongly positively invariant.

The set  $A$  will be called 'weakly stable' if for any  $\varepsilon > 0$  there is a  $\delta > 0$  such that for any point  $x \in S_\delta(A)$  there is a trajectory  $\phi(x, t)$  satisfying  $\phi(x, R^+) \subset S_\varepsilon(A)$ .

The set  $A$  will be called 'strongly quasi-asymptotically stable' if for some fixed  $\delta > 0$  and any given  $\varepsilon > 0$  there is a  $T(\varepsilon) \geq 0$  such that  $x \in S_\delta(A) \Rightarrow F(x, t) \subset S_\varepsilon(A)$  for all  $t \geq T(\varepsilon)$ .

The set  $A$  will be called 'weakly quasi-asymptotically stable' if for some fixed  $\delta > 0$  and any given  $\varepsilon > 0$  there is a  $T(\varepsilon) \geq 0$  such that through any point  $x \in S_\delta(A)$  there is a trajectory satisfying  $\phi(x, t) \subset S_\varepsilon(A)$  for all  $t \geq T(\varepsilon)$ , and  $\phi(x, t)$  can be chosen independently of  $\varepsilon$ .

The set  $A$  will be called 'strongly asymptotically stable' if it is strongly stable and strongly quasi-asymptotically stable.

The set  $A$  will be called 'weakly asymptotically stable' if it is weakly stable and weakly quasi-asymptotically stable.

The set  $A$  will be called 'strongly quasi-asymptotically stable in the large' if for any  $\alpha > 0$  and any  $\varepsilon > 0$  there is a  $\beta(\alpha) > 0$  and a  $T(\alpha, \varepsilon) \geq 0$  such that  $x \in S_\alpha(A) \Rightarrow F(x, t) \subset S_\beta(A)$  for all  $t \geq 0$  and  $F(x, t) \subset S_\varepsilon(A)$  for all  $t \geq T(\alpha, \varepsilon)$ .

The set  $A$  will be called 'weakly quasi-asymptotically stable in the large' if for any point  $x$  there is a certain trajectory  $\phi(x, t)$  such that for any  $\alpha > 0$  and any  $\varepsilon > 0$  there is a  $\beta(\alpha) > 0$  and a  $T(\alpha, \varepsilon) \geq 0$  satisfying  $x \in S_\alpha(A) \Rightarrow \phi(x, t) \subset S_\beta(A)$  for all  $t \geq 0$  and  $\phi(x, t) \subset S_\varepsilon(A)$  for all  $t \geq T(\alpha, \varepsilon)$ .

The set  $A$  will be called 'strongly asymptotically stable in the large' if it is strongly stable and strongly quasi-asymptotically stable in the large.

The set  $A$  will be called 'weakly asymptotically stable in the large' if it is weakly stable and weakly quasi-asymptotically stable in the large.

These definitions are the straightforward translations of the classical definitions to control theory, as explained before. In a similar way, all refinements of the concepts of stability, e.g. for non-autonomous systems, can be translated to control theory in a strong and a weak form.

### Characterization of Invariance and Stability by Means of Liapunov's Function

Similar to the case of ordinary dynamical systems, the invariance and stability properties of sets can be characterized by Liapunov-type functions. Theorems on the strong form of stability were already given by Zubov<sup>9</sup>, the parallel development of theorems concerning the weak form can be accomplished without major difficulties. Sometimes the proofs are more subtle because of the fact that if for some sequence  $0 < t_1 < t_2 < t_3 \dots$  one has  $x_i \in F(x_0, t_i)$ , it is not possible to deduce from this that there exists a trajectory passing through all  $x_i$  (this is not in general the case). Nevertheless, the following theorems are given without proof as examples of the power of the second method of Liapunov applied to these problems; before stating them some definitions must be given.

If  $A$  is a given set, it is said that the scalar function  $V(x)$  is *positive definite* ( $A$ ) if:

- (i)  $V(x)$  is lower semicontinuous [this is,  $\lim_{y \rightarrow x} V(y) \geq V(x)$ ].
- (ii)  $V(x) \leq 0$  for  $x \in A$ ,  $V(x) > 0$  for  $x \notin A$ .
- (iii) There exist two continuous monotonically increasing functions  $v(r)$  and  $w(r)$  of the real variable  $r \geq 0$ , such that  $v(0) = w(0) = 0$  and  $v[\rho(x, A)] \leq V(x) \leq w[\rho(x, A)]$ .

From this definition it follows that the set  $A$  is closed.

One calls generalized upper total derivative the expression

$$\dot{V}^+(x) = \lim_{t \rightarrow 0^+} \text{l.u.b.} \left\{ \frac{V[F(x, \tau)] - V(x)}{\tau}; 0 < \tau \leq t \right\}$$

Similarly the generalized lower total derivative is the expression

$$\dot{V}^-(x) = \lim_{t \rightarrow 0^+} \text{g.l.b.} \left\{ \frac{V[F(x, \tau)] - V(x)}{\tau}; 0 < \tau \leq t \right\}$$

The generalized total derivatives  $\bar{V}^-(x)$  and  $\dot{V}^-(x)$  are defined in the same way with  $t \rightarrow 0^-$ ,  $0 > \tau \geq t$ . The generalized total derivatives  $\bar{V}(x)$  and  $\dot{V}(x)$  correspond to  $t \rightarrow 0$ ,  $-t \leq \tau \leq t$ .

Now we are able to state the following theorems.

**Theorem 1:** If  $V(x)$  is any function such that  $\bar{V}^+(x) \leq 0$ , then the set  $A(\lambda) = \{x; V(x) \leq \lambda\}$ , supposed non-empty, is a strongly positively invariant set.

**Theorem 2:** If  $V(x)$  is lower semicontinuous and  $\dot{V}^+(x) \leq 0$ , then the set  $A(\lambda) = \{x; V(x) \leq \lambda\}$ , supposed non-empty, is a weakly positively invariant set.

**Theorem 3:** If  $A$  is a closed set,  $V(x)$  is positive definite ( $A$ ), and  $\bar{V}^+(x) \leq 0$ , then  $A$  is strongly stable.

**Theorem 4:** If  $A$  is a closed set,  $V(x)$  is positive definite ( $A$ ) and  $\dot{V}^+(x) \leq 0$ , then  $A$  is weakly stable.

**Theorem 5:** If  $A$  is a closed set,  $V(x)$  is positive definite ( $A$ ) and  $\bar{V}^+(x)$  is negative definite ( $A$ ), then  $A$  is strongly asymptotically stable.

**Theorem 6:** If  $A$  is a closed set,  $V(x)$  is positive definite ( $A$ ) and  $\dot{V}^+(x)$  is negative definite ( $A$ ), then  $A$  is weakly asymptotically stable.

In all these theorems the function  $V(x)$  is supposed to be defined in a suitable neighbourhood of the set  $A$ . The importance of the theorems is due to the physical meaning of the given definitions and to the fact that the upper and lower generalized total derivatives can be (as a matter of principle) evaluated without any reference to trajectories or possible choices of control functions. In the definitions of those derivatives, only the behaviour of the  $V(x)$  functions of the attainable part of some neighbourhood of  $x_0$  plays any role. In very complicated systems, for example biological or social systems, the lack of knowledge of the equations governing the motion, or of the kind of control, does not necessarily directly concern this approach.

#### Relation with Classical Results

If one wishes to introduce differentiability conditions which approach the classical cases, one considers the space  $X$  to be

Euclidean  $n$ -dimensional; to make matters more appealing to the geometrical intuition it is supposed that the  $n$ th coordinate  $x^n$  represents the time  $t$  (as is usually done for passing from non-autonomous to autonomous systems). The expression

$$C(x_0) = \lim_{t \rightarrow 0^+} \frac{F(x_0, t) - x_0}{t}$$

has to be understood as a limit of a variable set in the Hausdorff metric for subsets of  $X$ . If this limit exists and is compact, it represents a cone with vertex in  $x_0$ , defining the tangents to all possible trajectories. More generally, a trajectory  $x(t)$  will be admissible if for any sequence  $t_i \rightarrow t_0^+$  ( $i = 1, 2, 3, \dots$ ) such that  $x_i = x(t_i) \rightarrow x_0$  and there exists the limit  $x_i - x_0 / t_i - t_0$ , this limit belongs to the cone  $C(x_0)$ . But this is exactly the definition of an equation in contingents<sup>5</sup>. The definition can be slightly generalized in the way that limits of the quotient of the type  $x_i - x_j / t_i - t_j$  are considered and then it is called an equation in paratingents<sup>8</sup>. These were considered by Zaremba and Marchaud, as already mentioned. Under some assumptions about continuity and convexity, a purely geometrical approach to Pontrjagin's maximum principle in the theory of optimal control can be made<sup>7</sup>. By linking all these intermediate steps together, one can expect to obtain many known results in a purely axiomatic way. Of course, this will be of real advantage only if the assumptions made are really more general, but if only the results obtained appear more naturally and clearer, this approach will also be useful.

Finally, mention must be made of the interest which might result in the study of systems which do not satisfy all axioms. For example, if axiom (IV) is omitted, systems are included where some states are not the result of the evolution from some previous state, but give a starting point for a new evolution. The author is not in a position, and it would not be in the scope of this paper, to discuss the possibility of application of such systems.

#### References

- <sup>1</sup> BARBASHIN, E. A. On the theory of generalized dynamical system, *Uch. Zap. M.G.U.* No. 135 (1949) 110-133
- <sup>2</sup> BIRKHOFF, G. D. Dynamical systems, *A.M.S. Publ.* (1927, repr. 1952)
- <sup>3</sup> GOTTSCHALK, W. H., and HEDLUND, G. A. Topological dynamics, *A.M.S. Publ.* (1955)
- <sup>4</sup> KELLEY, J. L. Hyperspaces of a continuum, *Trans. A.M.S.* 52 (1942), pp. 22-36
- <sup>5</sup> MARCHAUD, A. Sur les champs continus de demi-cones convexes et leurs integrales, *Compositio Math.* 3 (1936), pp. 89-127
- <sup>6</sup> MICHAEL, E. Topologies on spaces of subsets, *Trans. A.M.S.* 71 (1951), pp. 152-182
- <sup>7</sup> ROXIN, E. A geometric interpretation of Pontrjagin's maximum principle, *RIAS Rep.* 61-15 (1961)
- <sup>8</sup> ZAREMBA, S. C. Sur les equations au paratingent, *Bull. Sci. Math.* 60 (1936), pp. 139-160
- <sup>9</sup> ZUBOV, V. I. Methods of A. M. Liapunov and their application, *Izdat. Leningr. Univ.* (1957) AEC-Transl.; 4439, *US. AEC.*

## DISCUSSION

**Author's Opening Remarks**

The following are three recent publications about, or related to, the subject of this paper:

(1) Roxin: *R.I.A.S. Tech. Rep. 62-16*. In this work all the proofs mentioned in the paper are given for the more general case of non-autonomous systems.

(2) Wazewski: Several papers in the *Bull. Acad. Polonaise de Sciences*, since 1960, developing the theory of contingent equations (Marchaud-Zaremba).

(3) Bushaw: *R.I.A.S. Tech. Rep. 63-10*. 'Dynamical polysystems and optimization', where the basic element defined axiomatically is the trajectory.

It would be interesting to establish the conditions under which these different approaches are equivalent to one another.

W. DE BACKER, *Cetis Euratom CCR, Ispra, Italy*

Supposing that a set of differential equations describes a dynamical system, then the same can be said of the corresponding attainability functions. Between both descriptions there exists a mathematical

relationship; this relationship, however, can be very complicated as far as the computational aspects using electronic computers are concerned. This means that it would not be the same for an electronic computer if some problems were stated in terms of differential equations or in terms of attainability functions. In putting forward this remark I have in mind a very important class of optimization problems, namely those concerning economical systems. I think, although of course it is only a matter of intuition, that most of the present economic optimizers, who are individuals, governments and planning teams, have much better estimates of the attainability functions than of the differential equations of the economic dynamical system they are faced with.

E. ROXIN, *in reply*

Indeed, mathematical generalizations and axiomatic foundations of theories are not only made for elegance of presentation, but mainly in order to be able to apply them to cases not included in the original theory. Economical, social and biological systems, where the differential equation governing its behaviour is extremely difficult to know and even the existence of such an equation is doubtful, might be good examples for application of the present theory.

# The Inverse Problem of Integral Square Estimation of Transient Responses

W. JAROMINEK

## Summary

The inverse problem of integral square estimation may be treated as a problem of inversion of the transient performance of linear stationary systems. The present paper discusses a method of solving this problem by means of spectra of square integral estimation introduced for this purpose. The analytical expressions obtained allow one to determine in a unique manner the transfer functions corresponding to any integral square estimation given by means of these spectra. They are therefore particularly useful for the synthesis of simple and multiple stationary linear systems; at the same time they enable one to select in a simple way the values of the parameters to satisfy the required static and dynamic characteristics of the systems.

## Sommaire

Le problème inverse de l'estimation de l'intégrale du carré de l'écart peut être considéré comme un problème d'inversion du régime transitoire des systèmes linéaires stationnaires. Dans le présent travail on propose une méthode de solution de ce problème au moyen de spectres de l'intégrale du carré de l'écart introduits spécialement dans ce but. Les expressions analytiques obtenues permettent de déterminer d'une façon unique les transmittances correspondant aux valeurs quelconques, données a priori, de l'intégrale du carré de l'écart données au moyen de ces spectres. Par conséquent ces expressions sont particulièrement utiles pour la synthèse des systèmes linéaires stationnaires asservis simples ou multiples. De même elles donnent la possibilité de choisir d'une manière suffisamment simple les valeurs des paramètres satisfaisant aux caractéristiques exigées, statiques et dynamiques, du système.

## Zusammenfassung

Das inverse Problem zur Bestimmung des quadratischen Integralkriteriums läßt sich als umgekehrte Aufgabe zur Gütebestimmung des Übergangsverhaltens in linearen stationären Systemen betrachten. Die Arbeit erläutert die Lösungsmethode dieser Aufgabe; sie stützt sich auf speziell für diese Zwecke eingeführte Spektren des quadratischen Integralkriteriums. Die erhaltenen analytischen Beziehungen ermöglichen es, die Übertragungsfunktionen entsprechend den Spektren des quadratischen Integralkriteriums eindeutig zu bestimmen. Deshalb eignen sich diese Beziehungen besonders zur Synthese von ein- und mehrfachen linearen stationären Systemen. Die Spektren des quadratischen Integralkriteriums erlauben es gleichzeitig in einfacher Weise die Parameterwerte, die die statischen und dynamischen Charakteristiken des Regelungssystems erfüllen, zu bestimmen.

## Introduction

This paper is a continuation of the author's work devoted to the investigation of automatic control systems by means of determinant indices of stability margin<sup>1</sup>. One of the principal problems considered in that reference is the inverse stability problem of linear systems; that is, the problem of obtaining

expressions enabling the characteristic equation for prescribed values of the indices of stability margin to be established.

The paper is devoted to the inverse of integral square estimation of transient responses. The inverse integral square estimation may be considered to constitute the inverse problem of quality of a transient process in a linear system. The solution of this problem is obtained by introducing the notion of spectrum of the integral square estimation. The expressions obtained enables the determination, in a unique form, of the transfer functions corresponding to a prescribed qualitative evaluation according to the integral square estimation of transient responses, thus being particularly useful for the synthesis of one- and multi-loop systems.

This work was done under the direction of the Academician B. N. Petrov to whom the author wishes to express his gratitude for many valuable remarks and suggestions.

## 1. The Inverse Stability Problem of Linear Systems

The starting point of the present paper is the inverse stability problem of linear systems. The integral square estimation expressed in the form proposed by A. Krasovskii and called, in what follows, the Krasovskii integral criterion or the Krasovskii evaluation<sup>2, 3</sup>, is the assumed estimation of quality. For the sake of comparability of results the normalized Krasovskii evaluations  $\bar{J}_n^{(m)}$  are considered. One has

$$\bar{J}_n^{(m)} = \bar{J}_n^{(m)}[\bar{F}(p)] \quad (1)$$

$$\bar{F}(p) = \frac{\bar{B}(p)}{\bar{A}(p)} = \frac{b_0 p^m + b_1 p^{m-1} + \dots + b_{m-1} p + 1}{p^n + a_1 p^{n-1} + \dots + a_{n-1} p + 1} \quad (2)$$

where  $\bar{F}(p)$  is the normalized transfer function and  $n > m \geq 0$ .

The consideration of normalized Krasovskii evaluation and normalized transfer functions  $\bar{F}(p)$  does not affect the generality of the assumptions.

It has been shown<sup>1, 4, 5</sup> that the Markov stability criterion enables a solution of the inverse stability problem of linear systems to be obtained. The generalized notion of determinant indices of stability margin has also been introduced<sup>1, 4, 5</sup>; the indices will be denoted by SMI (Stability Margin Indices).

The determination of the values of the coefficients of the characteristic equation corresponding to arbitrary values of the SMI is obtained according to the developed method<sup>1</sup> by intermediate determination of the Markov parameters. To omit the intermediate stage (the determination and calculation of Markov parameters), which is specially convenient in the case of synthesis of linear systems based on the qualitative Krasovskii's integral criterion, a new method has been developed for establish-

ing characteristic equations, corresponding to any prescribed conditions concerning the *SMI*<sup>6, 7</sup>. It presents a new and independent solution of the inverse stability problem.

## 2. Expansion of the Coefficients of the Characteristic Equation in Terms of *SMI*

The new solution of the inverse stability problem consists in expansion of the coefficients of the characteristic equation in terms of determinant indices of stability margin and, in particular cases, in terms of the Hurwitz or Markov determinants or Routh parameters. As an example of the expansion of the coefficients of the normalized characteristic equation in terms of Hurwitz determinants, mention should be made of *Table 2*, Reference 1. To generalize the results obtained there to the case of any degree 'n' write the characteristic equation  $A_n(p) = 0$  in the following form<sup>7</sup>:

$$A_n(p) = p^n + a_{1,n} p^{n-1} + a_{2,n} p^{n-2} + \dots + a_{k,n} p^{n-k} + \dots + a_{n,n} \quad (3)$$

By considering the sequence of Routh's matrices corresponding to successive values of the degree  $n$  and the equivalent sequence of Hurwitz matrices, it can be shown that the coefficients  $a_{k,n}$  ( $k = 1, 2, \dots, n$ ) of the characteristic polynomial (3) can be expressed in a unique form in terms of Hurwitz determinants<sup>7</sup>. In particular, the following expansions of the coefficients  $a_{k,n}$  are obtained in terms of Hurwitz determinants  $\Delta_{k,n} \equiv \Delta_k$ :

$$\left. \begin{aligned} a_{1,n} &= \frac{\Delta_1}{\Delta_0} \\ a_{2,n} &= \frac{\Delta_2}{\Delta_1} + \sum_{i=1}^{n-2} \frac{\Delta_{i-1}}{\Delta_i} \cdot \frac{\Delta_{i+2}}{\Delta_{i+1}} \\ a_{3,n} &= \frac{\Delta_3}{\Delta_1} + \frac{\Delta_1}{\Delta_0} \sum_{i=2}^{n-2} \frac{\Delta_{i-1}}{\Delta_i} \cdot \frac{\Delta_{i+2}}{\Delta_{i+1}} \end{aligned} \right\} \quad (4)$$

$$a_{2k-1, 2k} = \frac{\Delta_{2k-1}}{\Delta_{2k-2}} + \frac{\Delta_{2k}}{\Delta_{2k-1}} \cdot \sum_{i=1}^{k-1} \frac{\Delta_{2i-1}}{\Delta_{2i-2}} \cdot \frac{\Delta_{2i-1}}{\Delta_{2i}} \quad (5)$$

$$a_{2k, 2k+1} = \frac{\Delta_{2k}}{\Delta_{2k-1}} + \frac{\Delta_{2k+1}}{\Delta_{2k}} \left( \frac{\Delta_0}{\Delta_1} + \sum_{i=1}^{k-1} \frac{\Delta_{2i}}{\Delta_{2i-1}} \cdot \frac{\Delta_{2i}}{\Delta_{2i+1}} \right) \quad (6)$$

$$a_{n,n} = \frac{\Delta_n}{\Delta_{n-1}} \quad (7)$$

*Table 1* has been prepared on the basis of eqns (4)–(7).

In the general case the expansion of the coefficients in terms of Hurwitz determinants is expressed by the following algorithm<sup>7</sup>:

$$a_{k,n} = a_{k,n-1} + \frac{\Delta_{n-3} \Delta_n}{\Delta_{n-2} \Delta_{n-1}} a_{k-2, n-2} \quad (8)$$

where  $a_{i,s} \equiv 1$  in the case of  $i = 0$  and  $0 \leq s \leq n$   
 $a_{\alpha,\alpha} \equiv 1$  in the case of  $\alpha = 0, -1, -2, \dots$   
 $a_{i,s} \equiv 0$  in the case of  $i < 0$  or  $i > s$

The expressions for the expansion of the coefficients  $a_{k,n}$  can be most easily obtained by means of the recurrence equations

$$A_k(p) = p A_{k-1}(p) + \frac{\Delta_{k-3} \Delta_k}{\Delta_{k-2} \Delta_{k-1}} A_{k-2}(p) \quad (9)$$

where  $A_k(p)$  is a polynomial of degree  $k$  and  $\Delta_l \equiv 1$  for  $l = 0, -1, -2, \dots$ . The recurrence equation (9) holds for  $1 \leq k \leq n$  (Reference 7). In particular, in the case of  $k = 1, 2$  one obtains  $A_0(p) = A_{-1}(p) \equiv 1$ .

Analogous expressions may be derived for the remaining forms of the *SMI*<sup>7</sup>.

Equations (4)–(9) enable the inverse stability problem of linear systems to be easily solved. The selection of appropriate values of the *SMI* should be done on the basis of a suitable qualitative criterion of transient responses.

## 3. The Transformed Krasovskii Integral Criterion

As an estimation of quality of transient responses assume the Krasovskii integral criterion  $\bar{J}_n^{(m)}$ . It has been shown<sup>1, 5, 8</sup> that as a result of a suitable transformation the integral square estimation  $\bar{J}_n^{(0)}$  can be expressed in a simple manner in terms of the indices of stability margin. The transformed evaluation  $\bar{J}_n^{(0)}$  takes, when Hurwitz determinants are used, the form

$$\bar{J}_{2k}^{(0)} = \frac{1}{2} \left( \frac{a_{2k-1}}{a_{2k}} + \frac{a_0}{a_1} + \frac{\Delta_2 \Delta_2}{\Delta_1 \Delta_3} + \frac{\Delta_4 \Delta_4}{\Delta_3 \Delta_5} + \dots + \frac{\Delta_{2(k-1)}}{\Delta_{2k-3}} \cdot \frac{\Delta_{2(k-1)}}{\Delta_{2k-1}} \right) \quad (10)$$

$$= \frac{1}{2} \left( \frac{a_{2k-1}}{a_{2k}} + \frac{a_0}{a_1} + \sum_{i=2}^k \frac{\Delta_{2(i-1)}}{\Delta_{2i-3}} \cdot \frac{\Delta_{2(i-1)}}{\Delta_{2i-1}} \right)$$

$$\begin{aligned} \bar{J}_{2k+1}^{(0)} &= \frac{1}{2} \left( \frac{a_{2k}}{a_{2k+1}} + \frac{\Delta_1 \Delta_1}{\Delta_0 \Delta_2} + \frac{\Delta_3 \Delta_3}{\Delta_2 \Delta_4} + \frac{\Delta_5 \Delta_5}{\Delta_4 \Delta_6} + \dots + \frac{\Delta_{2k-1}}{\Delta_{2k-2}} \cdot \frac{\Delta_{2k-1}}{\Delta_{2k}} \right) \\ &= \frac{1}{2} \left( \frac{a_{2k}}{a_{2k+1}} + \sum_{i=1}^k \frac{\Delta_{2i-1}}{\Delta_{2i-2}} \cdot \frac{\Delta_{2i-1}}{\Delta_{2i}} \right) \end{aligned} \quad (11)$$

In order to obtain a complete transformation of the evaluation  $\bar{J}_n^{(0)}$  make use of the relations (5)–(7) and express the ratios  $a_n - 1/a_n$  for the coefficients of the characteristic equation  $A_n(p)$  in terms of Hurwitz determinants. It is found that

$$\frac{a_{2k-1}}{a_{2k}} = \sum_{i=1}^k \frac{\Delta_{2i-1}}{\Delta_{2i-2}} \cdot \frac{\Delta_{2i-1}}{\Delta_{2i}} \quad (12)$$

$$\frac{a_{2k}}{a_{2k+1}} = \frac{\Delta_0}{\Delta_1} + \sum_{i=1}^k \frac{\Delta_{2i}}{\Delta_{2i-1}} \cdot \frac{\Delta_{2i}}{\Delta_{2i+1}} \quad (13)$$

Observe that the evaluations (10) and (11) have different forms in case of even ( $n = 2k$ ) and odd ( $n = 2k + 1$ ) degrees  $n$  of the characteristic equation. Substituting in (10) the expression (12) and in (11) the expression (13) and performing an appropriate change of summation indices one obtains a single general expression

$$\bar{J}_n^{(0)} = \frac{1}{2} \sum_{i=1}^n \frac{\Delta_{i-1}}{\Delta_{i-2}} \cdot \frac{\Delta_{i-1}}{\Delta_i} \quad (14)$$

where  $\Delta_0 \equiv \Delta_{-1} \equiv 1$  is arbitrarily assumed.

The expression (14) is the transformed Krasovskii evaluation expressed exclusively in terms of indices of stability margin\*.

\* Other forms of the transformed Krasovskii evaluation may be found in Reference 7.

Table 1. Expansion of the coefficients of the characteristic equation in terms of indices of stability margin of the group  $H$  (the general case)

$n$	Coefficients of the characteristic equation					
	$a_0$	$a_1$	$a_2$	$a_3$	$a_4$	$a_5$
1	1	$a_1 = \frac{\Delta_1}{\Delta_0}$				
2	1	$a_1 = \frac{\Delta_1}{\Delta_0}$	$a_2 = \frac{\Delta_2}{\Delta_1}$			
3	1	$a_1 = \frac{\Delta_1}{\Delta_0}$	$a_2 = \frac{\Delta_2}{\Delta_1} + \frac{\Delta_0 \Delta_3}{\Delta_1 \Delta_2}$	$a_3 = \frac{\Delta_3}{\Delta_2}$		
4	1	$a_1 = \frac{\Delta_1}{\Delta_0}$	$a_2 = \frac{\Delta_2}{\Delta_1} + \frac{\Delta_0 \Delta_3}{\Delta_1 \Delta_2} + \frac{\Delta_1 \Delta_4}{\Delta_2 \Delta_3}$	$a_3 = \frac{\Delta_3}{\Delta_2} + \frac{\Delta_1}{\Delta_0} \cdot \frac{\Delta_1 \Delta_4}{\Delta_2 \Delta_3}$	$a_4 = \frac{\Delta_4}{\Delta_3}$	
5	1	$a_1 = \frac{\Delta_1}{\Delta_0}$	$a_2 = \frac{\Delta_2}{\Delta_1} + \frac{\Delta_0 \Delta_3}{\Delta_1 \Delta_2} + \frac{\Delta_1 \Delta_4}{\Delta_2 \Delta_3} + \frac{\Delta_2 \Delta_5}{\Delta_3 \Delta_4}$	$a_3 = \frac{\Delta_3}{\Delta_2} + \frac{\Delta_1}{\Delta_0} \left( \frac{\Delta_1 \Delta_4}{\Delta_2 \Delta_3} + \frac{\Delta_2 \Delta_5}{\Delta_3 \Delta_4} \right)$	$a_4 = \frac{\Delta_4}{\Delta_3} + \frac{\Delta_5}{\Delta_4} \left( \frac{\Delta_0}{\Delta_1} + \frac{\Delta_2 \Delta_2}{\Delta_1 \Delta_3} \right)$	$a_5 = \frac{\Delta_5}{\Delta_4}$
6	1	$a_1 = \frac{\Delta_1}{\Delta_0}$	$a_2 = \frac{\Delta_2}{\Delta_1} + \frac{\Delta_0 \Delta_3}{\Delta_1 \Delta_2} + \frac{\Delta_1 \Delta_4}{\Delta_2 \Delta_3} + \frac{\Delta_2 \Delta_5}{\Delta_3 \Delta_4} + \frac{\Delta_3 \Delta_6}{\Delta_4 \Delta_5}$	$a_3 = \frac{\Delta_3}{\Delta_2} + \frac{\Delta_1}{\Delta_0} \left( \frac{\Delta_1 \Delta_4}{\Delta_2 \Delta_3} + \frac{\Delta_2 \Delta_5}{\Delta_3 \Delta_4} + \frac{\Delta_3 \Delta_6}{\Delta_4 \Delta_5} \right)$	$a_4 = \frac{\Delta_4}{\Delta_3} + \frac{\Delta_5}{\Delta_4} \left( \frac{\Delta_0}{\Delta_1} + \frac{\Delta_2 \Delta_2}{\Delta_1 \Delta_3} \right) + \frac{\Delta_5}{\Delta_4} \left( \frac{\Delta_0}{\Delta_1} + \frac{\Delta_2 \Delta_2}{\Delta_1 \Delta_3} \right)$	$a_5 = \frac{\Delta_5}{\Delta_4} + \frac{\Delta_6}{\Delta_5} \left( \frac{\Delta_1 \Delta_1}{\Delta_0 \Delta_2} + \frac{\Delta_3 \Delta_3}{\Delta_2 \Delta_4} \right)$

Notes: (1) In the case of normalized characteristic equation one should substitute  $a_n = \frac{\Delta_n}{\Delta_{n-1}} \equiv 1$ .

(2)  $n$  is the degree of the characteristic equation.

(3)  $\Delta_0 \equiv 1$  (assumed arbitrarily).

It holds for both even and odd degrees; that is, for any degree  $n$  of the characteristic equation.

The above transformation of the Krasovskii evaluation may be considered as a transition from one set of independent variables to another. The independent variables of the first set are the coefficients of transfer function; those of the other, the *SMI*. Further investigations show that this transformation is of essential importance chiefly because the *SMI* supply much more necessary information on the control system than the transfer function coefficients. It is also of importance that the new expressions of the integral square estimation take a much simpler analytical form, which is essential for the synthesis of control systems.

To generalize the results obtained to systems of the non-zero class ( $m \neq 0$ ;  $n > m > 0$ ) consider some of the relations between the Krasovskii determinants and the *SMI*.

#### 4. Expansion of the Krasovskii Determinants in Terms of *SMI*

In the general case the normalized<sup>†</sup> integral square estimation takes the form

$$\bar{J}_n^{(m)} = \frac{1}{2} \sum_{\alpha=0}^m B_{m-\alpha}^{(m)} \cdot \frac{\Delta_{m-\alpha}^{(n)}}{\Delta_n} - b_{m-1} \quad (15)$$

where

$$B_{m-\alpha}^{(m)} = b_{m-\alpha}^2 - 2b_{m-\alpha+1}b_{m-\alpha-1} + 2b_{m-\alpha+2}b_{m-\alpha-2} + \dots + 2(-1)^{m-\alpha}b_m b_{m-2\alpha} \quad (16)$$

for  $\alpha = 0, 1, 2, \dots, m$  and  $b_m \equiv 1$ ;  $b_k \equiv 0$  ( $k < 0$ ;  $k > m$ ).

The expression of the normalized evaluation (15) in terms of *SMI* requires, above all, the expansion of the Krasovskii determinants  $\Delta_{m-\alpha}^{(n)}/\Delta_n$  in terms of *SMI*<sup>7, 8</sup>. The elements of these determinants are exclusively the coefficients of the characteristic equation, therefore the unique expansion  $\Delta_{m-\alpha}^{(n)}/\Delta_n$  in terms of *SMI* may be done on the basis of eqns (4)–(9). As an example a few of the relations obtained are quoted:

$$\frac{\Delta_m^{(n)}}{\Delta_n} = \sum_{i=1}^n \frac{\Delta_{i-1}}{\Delta_{i-2}} \cdot \frac{\Delta_{i-1}}{\Delta_i} \quad (17)$$

$$\frac{\Delta_{m-1}^{(n)}}{\Delta_n} = \frac{\Delta_{n-2}}{\Delta_{n-1}} \quad (18)$$

$$\frac{\Delta_{m-2}^{(n)}}{\Delta_n} = \frac{\Delta_{n-3}}{\Delta_{n-1}} \quad (19)$$

$$\frac{\Delta_{m-3}^{(n)}}{\Delta_n} = \frac{\Delta_{n-4}}{\Delta_{n-2}} + \frac{\Delta_{n-3}}{\Delta_{n-2}} \frac{\Delta_{n-3}}{\Delta_{n-1}} \quad (20)$$

.....

The expansions of the remaining expressions of the form

$$\frac{\Delta_{m-\alpha}^{(n)}}{\Delta_n}$$

<sup>†</sup> The normalized evaluation  $\bar{J}_n^{(m)}$  corresponds to the normalized transfer function  $(\bar{F})p$ , for which one has  $a_0 = a_n = b_m \equiv 1$ .

can be represented in a similar manner. Consider the sequence

$$\left\{ \frac{\Delta_{m-\alpha}^{(n)}}{\Delta_n} \right\}$$

of these expressions:

$$\frac{\Delta_m^{(n)}}{\Delta_n}, \frac{\Delta_{m-1}^{(n)}}{\Delta_n}, \frac{\Delta_{m-2}^{(n)}}{\Delta_n}, \dots, \frac{\Delta_{m-\alpha}^{(n)}}{\Delta_n}, \dots, \frac{\Delta_{m-m}^{(n)}}{\Delta_n} \quad (21)$$

It can be shown that the structure of the equation obtained for the expansion of each particular expression

$$\frac{\Delta_{m-\alpha}^{(n)}}{\Delta_n},$$

in terms of the *SMI* is independent of the degrees  $n$  and  $m$  of the transfer function polynomials and depends only on the ordinal number  $\alpha$  in the sequence

$$\left\{ \frac{\Delta_{m-\alpha}^{(n)}}{\Delta_n} \right\}$$

This property is very useful for the generalization of the considered problem for the case of  $m > 0$ .

#### 5. The Optimum Integral Estimation $\bar{J}_n^{(m)}$ in the Sense of *SMI*

The value of the evaluation (15) depends on the distribution of poles and zeros of the transfer function (2). Assume that in the general case the distribution of the zeros is independent of that of the poles. Then, the coefficients  $B_{m-\alpha}^{(m)}$  are also independent of the coefficients of the characteristic equation and cannot, in general, be expressed in terms of *SMI*. For any assigned distribution of transfer function poles there exists only one distribution of zeros of the polynomial  $\bar{B}(p)$  in the numerator of the transfer function, which, for the given assumptions, corresponds to the minimum value of the evaluation  $\bar{J}_n^{(m)}$ . Such a distribution of zeros will be called, in what follows, optimum in relation to the *SMI*. The determination of the corresponding optimum polynomial  $\bar{B}(p) = \bar{B}(p)_{\text{opt}}$  will be called the optimization of the integral square estimation  $\bar{J}_n^{(m)}$  in the sense of *SMI*<sup>7</sup>.

The determination of the values of the coefficients of the optimum polynomial  $\bar{B}(p)_{\text{opt}}$  reduces to that of the extremum (minimum) value of a function of many independent variables. To do this one must equal to zero the partial derivatives of the evaluation  $\bar{J}_n^{(m)}$  with respect to the coefficients of the polynomial  $\bar{B}(p)$ . One has

$$\frac{\partial \bar{J}_n^{(m)}}{\partial b_0} = 0; \frac{\partial \bar{J}_n^{(m)}}{\partial b_1} = 0; \dots; \frac{\partial \bar{J}_n^{(m)}}{\partial b_k} = 0; \dots; \frac{\partial \bar{J}_n^{(m)}}{\partial b_{m-1}} = 0 \quad (22)$$

Solving successively for each  $m$  the system of  $m$  equations  $\partial \bar{J}_n^{(m)} / \partial b_k \equiv 0$  ( $k = 0, 1, 2, \dots, m-1$ ) the values of the coefficients of the polynomial  $\bar{B}(p)_{\text{opt}}$  will be obtained, i.e., the optimum in the sense of *SMI*. Thus, for instance, in the case of  $m = 3$  one has:

$$b_0 = \frac{\Delta_{n-3}}{\Delta_{n-4}}; b_1 = \frac{\Delta_{n-2}}{\Delta_{n-3}}; b_2 = \frac{\Delta_{n-1}}{\Delta_{n-2}} + \frac{\Delta_{n-3}}{\Delta_{n-4}} \frac{\Delta_{n-3}}{\Delta_{n-2}}; b_3 = 1 \quad (23)$$

Table 2 contains expressions for the coefficients of optimum polynomials  $\bar{B}(p)_{\text{opt}}$ , obtained as a result of solution of the set of eqns (22) for a few successive values of the degree  $m$  of the polynomial  $\bar{B}(p)$ .



Table 2. Coefficients  $b_i$  ( $i = 0, 1, 2, \dots, m$ ) of the polynomials  $B(p) = B(p)_{\text{opt}}$ , satisfying the optimum conditions in the sense of the indices of stability margin  $\left(q_i = \frac{\Delta_{i+1}}{\Delta_i}\right)$

$m$	Coefficients of the polynomial $B(p) = b_0 p^m + b_1 p^{m-1} + \dots + b_{m-1} p + b_m = B(p)_{\text{opt}}$					
	$b_0$	$b_1$	$b_2$	$b_3$	$b_4$	$b_5$
1	$q_{n-2}$	1				
2	$q_{n-3}$	$q_{n-2}$	1			
3	$q_{n-4}$	$q_{n-3}$	$q_{n-2} + \frac{q_{n-4}}{q_{n-3}}$	1		
4	$q_{n-5}$	$q_{n-4}$	$q_{n-3} + \frac{q_{n-5}}{q_{n-3}} + \frac{q_{n-5}}{q_{n-4}} \cdot q_{n-2}$	$q_{n-2} + \frac{q_{n-4}}{q_{n-3}}$	1	
5	$q_{n-5}$	$q_{n-5}$	$q_{n-4} + \frac{q_{n-6}}{q_{n-3}} + \frac{q_{n-6}}{q_{n-4}} \cdot q_{n-2} + \frac{q_{n-6}}{q_{n-5}} \cdot q_{n-3}$	$q_{n-3} + \frac{q_{n-5}}{q_{n-3}} + \frac{q_{n-5}}{q_{n-4}} \cdot q_{n-2}$	$q_{n-2} + \frac{q_{n-4}}{q_{n-3}} + \frac{q_{n-6}}{q_{n-5}}$	1

Notes: (1)  $n$  is the degree of the characteristic equation  $A(p) = 0$

(2)  $m$  is the degree of the polynomial  $B(p)$ ;  $0 \leq m < n$

(3)  $B(p)$  is the polynomial in the numerator of the normalized

$$\text{transfer function } \bar{F}(p) = \frac{A(p)}{B(p)}$$

The integral evaluation  $\bar{J}_n^{(m)}$ , that satisfies the set of conditions (22), will be called optimum in the sense of stability margin and denoted by  $\bar{J}_n^{(m)}_{\text{opt}}$ . Optimum evaluations in the sense of  $SMI$ , have a number of valuable properties. Some of them will be considered below. Of particular importance is the fact that for full analytical description of the evaluation  $\bar{J}_n^{(m)}_{\text{opt}}$  only the  $SMI$  are required.

## 6. The Two Equivalent Forms of the Integral Evaluation $\bar{J}_n^{(m)}$

In the general case the integral evaluation  $\bar{J}_n^{(m)}$  does not satisfy the optimum conditions (22), and therefore it cannot be expressed in terms of the  $SMI$  only. This follows directly from the assumption, that the coefficients of the polynomial  $\bar{B}(p)$  are independent of the coefficients of the characteristic equation  $\bar{A}(p)$ . In this connection try to separate in the integral evaluation  $\bar{J}_n^{(m)}$  a component depending exclusively on the  $SMI$  from another component in which the influence of the polynomial  $\bar{B}(p)$  is taken into account. The introduction of the  $SMI$  and the notion of optimum conditions in the sense of  $SMI$  enables two new equivalent forms of the integral evaluation  $\bar{J}_n^{(m)}$  to be established, that is:

$$\bar{J}_n^{(m)} = \bar{J}_n^{(0)} + M_n^{(m)} \quad (24)$$

and

$$\bar{J}_n^{(m)} = \bar{J}_n^{(m)}_{\text{opt}} + \Delta M_n^{(m)} \quad (25)$$

A detailed analysis of expressions (24) and (25) will be shown later. Now one is satisfied with the statement that for the determination of the first components, that is  $\bar{J}_n^{(0)}$  and  $\bar{J}_n^{(m)}_{\text{opt}}$ ,

only  $SMI$  are needed. To find the remaining components, that is  $M_n^{(m)}$  and  $\Delta M_n^{(m)}$ , the knowledge of the polynomial  $\bar{B}(p)$  is also needed. In particular, the component  $M_n^{(m)}$  expresses the increase of the evaluation  $\bar{J}_n^{(m)}$  due to the fact that the polynomial increase  $B(p) = \bar{B}(p) - b_m$  has been taken into consideration, and the component  $\Delta M_n^{(m)}$  is the increase due to the introduction of the polynomial  $\Delta B(p) = \bar{B}(p) - \bar{B}(p)_{\text{opt}}$  in the numerator of the transfer function  $\bar{F}(p)$ . For further investigation form (25) will be of particular use.

## 7. The Primary Spectrum of the Integral Evaluation $\bar{J}_n^{(m)}$

Under the term of primary spectrum of the integral evaluation  $\bar{J}_n^{(m)}$  one will understand the expression

$$R_n = R_n(r_1, r_2, r_3, \dots, r_n) \quad (26)$$

The elements of the spectrum  $R_n$  are  $r_1, r_2, \dots, r_n$ . They are related to the  $SMI$  by the formulae

$$r_i = \frac{q_{i-2}}{q_{i-1}} = \frac{\Delta_{i-1} \cdot \Delta_{i-1}}{\Delta_{i-2} \cdot \Delta_i} = \frac{S_{i-1}^* S_{i-1}^*}{S_{i-2}^* S_i^*}; \quad (i = 1, 2, \dots, n) \quad (27)$$

where  $q_i$  are the Routh parameters,  $\Delta_i$  the Hurwitz determinants, and  $S_i^*$  the Markov determinants ( $\Delta_i = a_1^i \cdot S_i^*$ ).

Example:

$$r_1 = \frac{\Delta_0}{\Delta_1}; r_2 = \frac{\Delta_1 \Delta_1}{\Delta_0 \Delta_2}; r_3 = \frac{\Delta_2 \Delta_2}{\Delta_1 \Delta_3}; \dots, r_n = \frac{\Delta_{n-1} \cdot \Delta_{n-1}}{\Delta_{n-2} \cdot \Delta_n}$$

Knowing the values of the elements  $r_1, r_2, \dots, r_n$  the values of the corresponding indices of stability margin can easily be determined:

Routh parameters  $q_s$

$$\frac{1}{q_{k-1}} = r_1 r_2, \dots, r_k = \prod_{i=1}^k r_i \quad (k=1, 2, \dots, n) \quad (28)$$

Hurwitz determinants  $\Delta_k$

$$\frac{1}{\Delta_k} = r_1^k r_2^{k-1} r_3^{k-2}, \dots, r_k = \prod_{\alpha=1}^k r_\alpha^{k+1-\alpha} \text{ for } k=1, 2, \dots, n \quad (29)$$

Markov determinants  $S_k^*$ ; ( $S_1^* \equiv 1$ )

$$\frac{1}{S_k^*} = r_2^{k-1} r_3^{k-2}, \dots, r_k = \prod_{\alpha=2}^k r_\alpha^{k+1-\alpha} \text{ for } k=2, 3, \dots, n \quad (30)$$

The primary spectrum  $R_n$  determines uniquely the first components of the forms (24) and (25) of the integral evaluation  $\bar{J}_n^{(m)}$ . In particular, by virtue of eqns (14) and (27), one can write at once

$$\bar{J}_n^{(0)} = \frac{1}{2} (r_1 + r_2 + \dots + r_n) = \frac{1}{2} \sum_{i=1}^n r_i \quad (31)$$

It can also be shown<sup>7</sup> that when the optimum conditions (22) are satisfied, the expression of the evaluation  $\bar{J}_n^{(m)}$  takes the following exceptionally simple form

$$\bar{J}_{n \text{ opt}}^{(m)} = \frac{1}{2} (r_1 + r_2 + \dots + r_{n-m}) = \frac{1}{2} \sum_{i=1}^{n-m} r_i \quad (32)$$

where  $0 \leq m < n$ .

From (32) it follows that evaluation  $\bar{J}_n^{(m)}$  depends only on the first  $n-m$  elements of the spectrum  $R_n$  and is invariant in relation to the remaining ones. Thus, the elements  $r_1, r_2, \dots, r_{n-m}$  will be called weight (influence) elements and the remaining ones, that is  $r_{n-(m-1)}, r_{n-(m-2)}, \dots, r_n$ , independent or free ones\*. Observe that although the independent elements of the spectrum  $R_n$  show no influence on the value of the evaluation  $\bar{J}_n^{(m)}$ , they influence the character of the transient response. This is a separate problem and is not dealt with in this paper.

On the basis of eqns (27)–(30) the stability of a control system can easily be analysed. From this analysis it follows that if a control system is stable, all the elements of the spectrum  $R_n$  are positive. If, in addition, the system is physically real, these elements are bounded. The spectrum  $R_n$ , of which all the elements are different from zero and positive will be called 'essentially positive'.

Another interesting property of the spectrum  $R_n$  is now shown. It is known that the stability margin of the system is greater for greater values of SMI<sup>3,4</sup>, Hurwitz determinants, for instance. This means that the stability margin is greater for smaller values of elements of the spectrum  $R_n$ .

On the basis of the above results and considerations the following cardinal properties of the spectrum  $R_n$  can be formulated:

**Property I:** In order that a linear control system with the characteristic equation  $A(p) = p^n + a_1 p^{n-1} + \dots + a_n = 0$  is stable and physically real it is necessary and sufficient that the primary spectrum  $R_n = R_n(r_1, r_2, \dots, r_n)$  of the evaluation  $\bar{J}_n^{(m)}$

\* In the case of normalized transfer function the condition  $\prod_{i=1}^n r_i \equiv 1$  should be satisfied.

of this system is essentially positive and bounded, that is  $0 < r_i < \infty$  for  $i = 1, 2, \dots, n$ .

**Property II:** The primary spectrum  $R_n$  characterizes the transient performance in a linear control system of the order  $n$  and class  $m$ , because the sum of its weight elements  $r_i$  ( $i = 1, 2, \dots, n-m$ ) determines the value of the evaluation  $\bar{J}_n^{(m)}$  satisfying the optimum conditions in the sense of stability margin (SMI), that is

$$\bar{J}_{n \text{ opt}}^{(m)} = \frac{1}{2} \sum_{i=1}^{n-m} r_i$$

for  $m = 0, 1, 2, \dots, n-1$ .

## 8. The Secondary Spectrum of the Integral Evaluation $\bar{J}_n^{(m)}$

The components  $M_n^{(m)}$  and  $\Delta M_n^{(m)}$  in expressions (24) and (25) for the evaluations  $\bar{J}_n^{(m)}$  depend in the general case on the spectrum  $R_n$  and the polynomial  $\bar{B}(p) = b_0 p^m + b_1 p^{m-1} + \dots + b_{m-1} p + b_m$  or the equivalent polynomial  $C(p) \equiv \bar{B}(p)$ , where

$$C(p) = c_m p^m + c_{m-1} p^{m-1} + \dots + c_1 p + c_0 \equiv \bar{B}(p); (c_0 = b_m = 1) \quad (33)$$

The task now is to find a set of  $m$  parameters such that their structure contains as much information as possible on the transient performance in a control system and would enable the determination, in a unique form, of the values of the coefficients  $c_\alpha = b_{m-\alpha}$  and the easy computation of the component  $M_n^{(m)}$  or  $\Delta M_n^{(m)}$  of  $\bar{J}_n^{(m)}$ . To this aim consider the partial derivatives

$$w_i = \frac{\partial M_n^{(m)}}{\partial c_i} = \sum_{\alpha=i}^m \frac{\partial M_\alpha}{\partial c_i} \quad \text{for } i = 1, 2, \dots, m \quad (34)$$

One has, in the case of odd  $i$ :

$$w_i = -1 + \sum_{l=1}^{E[\frac{m+1}{2}]} (-1)^{l+1} \cdot c_{2l-1} \cdot \frac{\Delta_{m-i}^{(n)}}{\Delta_n}$$

and

$$w_{2l+1} = \sum_{i=1}^{E[\frac{m+1}{2}]} (-1)^{(i+l+1)} \cdot c_{2i-1} \cdot \frac{\Delta_{m-(i+l)}}{\Delta_n} \quad (35)$$

where  $1 \leq l \leq E[m-1/2]$ . One finds, for even  $i$ :

$$w_{2l} = (-1)^l \cdot \frac{\Delta_{m-l}^{(n)}}{\Delta_n} + \sum_{i=1}^{E[\frac{m}{2}]} (-1)^{(i+l)} \cdot c_{2i} \cdot \frac{\Delta_{m-(i+l)}}{\Delta_n} \quad (36)$$

for  $1 \leq l \leq k = E[m/2]$ .

The expressions (35) constitute a set of equations for the odd coefficients  $c_\alpha$  of the polynomial (33) and the expressions (36) are a set of equations for even coefficients  $c_\alpha$  of that polynomial. The principal determinants of these systems will be denoted by  $W_m^{(1)}$  and  $W_m^{(2)}$  respectively, where

$$W_m^{(1)} = \begin{vmatrix} \frac{\partial w_1}{\partial c_1} & \frac{\partial w_1}{\partial c_3} & \dots \\ \frac{\partial w_3}{\partial c_1} & \frac{\partial w_3}{\partial c_3} & \dots \\ \dots & \dots & \dots \end{vmatrix} = \begin{vmatrix} w_{11} & w_{13} & w_{15} & \dots \\ w_{31} & w_{33} & w_{35} & \dots \\ w_{51} & w_{53} & w_{55} & \dots \\ \dots & \dots & \dots & \dots \end{vmatrix} \quad (37)$$

$$W_m^{(2)} = \begin{vmatrix} \frac{\partial w_2}{\partial c_2} & \frac{\partial w_2}{\partial c_4} & \dots \\ \frac{\partial w_4}{\partial c_2} & \frac{\partial w_4}{\partial c_4} & \dots \\ \dots & \dots & \dots \end{vmatrix} = \begin{vmatrix} w_{22} & w_{24} & w_{26} & \dots \\ w_{42} & w_{44} & w_{46} & \dots \\ w_{62} & w_{64} & w_{66} & \dots \\ \dots & \dots & \dots & \dots \end{vmatrix} \quad (38)$$

or

$$\left. \begin{aligned} W_m^{(1)} &= \frac{\partial(w_1, w_3, \dots, w_{2k-1}, \dots)}{\partial(c_1, c_3, \dots, c_{2k-1}, \dots)} = \frac{1}{q_{n-2} q_{n-3}, \dots, q_{n-k}}; \\ k &= 2E \left[ \frac{1+m}{2} \right] \\ W_m^{(2)} &= \frac{\partial(w_2, w_4, \dots, w_{2k}, \dots)}{\partial(c_2, c_4, \dots, c_{2k}, \dots)} = \frac{1}{q_{n-2} q_{n-3}, \dots, q_{n-l}}; \\ l &= 2E \left[ \frac{2+m}{2} \right] \end{aligned} \right\} \quad (39)$$

From eqns (37)–(39) it follows that the determinants  $W_m^{(1)}$  and  $W_m^{(2)}$  are the Jacobians of the transformation. The elements  $W_{ij}$  of these Jacobians are Krasovskii determinants

$$\frac{\Delta_{m-\alpha}^{(n)}}{\Delta_n}$$

with appropriate signs and

$$w_{ij} = \frac{\partial w_i}{\partial c_j} = \frac{\partial^2 M_n^{(m)}}{\partial c_i \partial c_j} \quad (i, j = 1, 2, 3, \dots, m) \quad (40)$$

Assume that the system is stable and its spectrum  $R_n$  is variable (constant); then, assume also that the Jacobians  $W_m^{(1)}$  and  $W_m^{(2)}$  have, in agreement with (39), constant values different from zero and positive. From the analysis it follows<sup>11</sup> that in this case all the necessary and sufficient conditions are satisfied for the transformation considered to be homeomorphic. It follows that the transformation of the set of parameters  $w_i$  ( $i = 1, 2, \dots, m$ ) in an  $m$ -dimensional space  $L^{(m)}$  into the set of parameters  $c_i$  ( $i = 1, 2, \dots, m$ ) in an  $m$ -dimensional space  $D^{(m)}$  is one-to-one, and that the homeomorphic representation of a space region is a space region and the representation of an arc is an arc. The set of the parameters  $w_i$  ( $i = 1, 2, \dots, m$ ) will be called the secondary spectrum of the integral evaluation  $\bar{J}_n^{(m)}$  and denoted by

$$w_m = w_m(w_1, w_2, \dots, w_m) \quad (41)$$

If the values of the elements  $w_i$  of the spectrum  $w_m$  are known, it is easy to calculate all the coefficients  $c_\alpha$  of the polynomial (33). To do this it suffices to solve in relation to  $c_\alpha$  the matrix equations:

$$\|W_m^{(1)}\| \cdot \|C_m^{(1)}\| = \|V_m^{(1)}\| \quad \text{and} \quad \|W_m^{(2)}\| \cdot \|C_m^{(2)}\| = \|V_m^{(2)}\| \quad (42)$$

The spectrum  $w_m$  is called positive, zero or negative if all its elements  $w_i$  ( $i = 1, 2, \dots, m$ ) are, respectively, positive or zero or negative. A spectrum  $w_m$  may also be of a mixed type. In particular, from the solution of eqns (42) it follows that if the spectrum  $w_m$  is zero, the set of eqns (22) is satisfied. This important feature of the spectrum  $w_m$  concerning the optimum evaluation  $\bar{J}_n^{(m)}$  in the sense of SMI can be expressed in the form of the following.

*Property of the spectrum  $w_m$ :* In order that the Krasovskii integral evaluation  $\bar{J}_n^{(m)}$  should satisfy the optimum conditions in the sense of stability margin (SMI) it is necessary and sufficient, that its secondary spectrum  $w_m$  is zero; that is,  $w_i \equiv 0$  ( $i = 1, 2, \dots, m$ ).

Now pass to another form of the spectrum  $w_m$  connected with the increment  $\Delta M_n^{(m)}$  of the evaluation  $\bar{J}_n^{(m)}$ . For this purpose the coefficients of the polynomial  $C(p)$  should first be represented in the form

$$c_i = c_{i \text{ opt}} + h_i \quad (i = 1, 2, \dots, m) \quad (43)$$

where  $c_{i \text{ opt}}$  satisfy the optimum conditions in the sense of SMI and expand  $M_n^{(m)}$  in Taylor's series for functions of more than one independent variable

$$\begin{aligned} M_n^{(m)}(c_{1 \text{ opt}} + h_1, c_{2 \text{ opt}} + h_2, \dots, c_{m \text{ opt}} + h_m) \\ = M_{n \text{ opt}}^{(m)} + \frac{dM_n^{(m)}}{1!} + \frac{d^2 M_n^{(m)}}{2!} + \dots + \frac{d^{k-1} M_n^{(m)}}{(k-1)!} + R_k \end{aligned} \quad (44)$$

In the general case the derivatives  $d^r M_n^{(m)}$  and the rest  $R_k$  of the expansion (44) are

$$d^v M_n^{(m)} = \left( \frac{\partial M_n^{(m)}}{\partial c_1} h_1 + \frac{\partial M_n^{(m)}}{\partial c_2} h_2 + \dots + \frac{\partial M_n^{(m)}}{\partial c_m} h_m \right)^v \quad (45)$$

and

$$R_k = \frac{d^k M_n^{(m)}}{k!} \quad (46)$$

where the derivatives  $d^r M_n^{(m)}$  for  $r < k$  should be determined at the point  $Q_{\text{opt}} = Q_{\text{opt}}(c_{1 \text{ opt}}, c_{2 \text{ opt}}, \dots, c_{m \text{ opt}})$  and the rest  $R_k$  at an intermediate point  $(C_{1 \text{ opt}} + \theta h_1, c_{2 \text{ opt}} + \theta h_2, \dots, c_{m \text{ opt}} + \theta h_m)$ , where  $0 < \theta < 1$ . One obtains

$$dM_n^{(m)} = \sum_{i=1}^m \frac{\partial M_n^{(m)}}{\partial c_i} h_i = \sum_{i=1}^m w_i h_i \quad (47)$$

$$R_k = R_2 = \frac{1}{2} \left( \sum_{i=1}^m \frac{\partial M_n^{(m)}}{\partial c_i} h_i \right)^2 = \frac{1}{2} \sum_{i=1}^m w_i h_i^2 + \sum_{\substack{i,j=1 \\ (i < j)}}^m w_{ij} h_i h_j \quad (48)$$

where

$$w_{ij} = \frac{\partial^2 M_n^{(m)}}{\partial c_i \partial c_j}$$

The expansion of the function  $M_n^{(m)}$  in Taylor's series, taking into account eqns (47) and (48), will now be written

$$M_n^{(m)} = M_{n \text{ opt}}^{(m)} + dM_n^{(m)} + R_2 \quad (49)$$

or

$$\Delta M_n^{(m)} = M_n^{(m)} - M_{n \text{ opt}}^{(m)} = dM_n^{(m)} + R_2$$

In agreement with the optimum theorem of the evaluation  $\bar{J}_n^{(m)}$  the point  $Q_{\text{opt}}(c_{1 \text{ opt}}, c_{2 \text{ opt}}, \dots, c_{m \text{ opt}})$  corresponds to a zero spectrum  $w_m$ . In other words the derivative (47) is at this point equal to zero, i.e.

$$dM_n^{(m)}(Q_{\text{opt}}) = 0 \quad (50)$$

It can easily be shown that the second partial derivatives  $w_{ij}$  do not depend on the choice of the intermediate point. Therefore the expression of the component  $\Delta M_n^{(m)}$  of the evaluation  $\bar{J}_n^{(m)}$  takes the following very simple form

$$\Delta M_n^{(m)} = R_2 = \frac{1}{2} \sum_{i=1}^m w_i h_i^2 + \sum_{\substack{i,j=1 \\ (i < j)}}^m w_{ij} h_i h_j \quad (w_{i, i+1} \equiv 0) \quad (51)$$

The partial derivatives  $w_{ii}$  and  $w_{ij}$  are Krasovskii determinants taken with an appropriate sign, therefore they depend on the spectrum  $R_n$  only. Analogous considerations show that the transformation of the set of parameters  $w_i$  into the set of parameters  $h_i$  ( $i = 1, 2, \dots, m$ ) is also homeomorphic; that is, one-to-one. In this connection the parameters  $h_i$  will be taken as elements of the second, equivalent form of the secondary spectrum  $w_m$ ; that is,

$$w_m(h_1, h_2, \dots, h_m) = w_m(w_1, w_2, \dots, w_m) \quad (52)$$

If the spectrum  $w_m$  is zero; that is,  $h_i \equiv 0$  ( $i = 1, 2, \dots, m$ ); then  $\Delta J_n^{(m)} = 0$  and  $\bar{J}_n^{(m)} = \bar{J}_n^{(m)}_{\text{opt}}$ .

The secondary spectrum  $w_m$  has a number of properties facilitating the qualitative analysis of the influence of distribution of zeros of the transfer function on the transient performance<sup>7, 12</sup>. Thus, for instance, a positive or negative spectrum  $w_m$  shows that the corresponding fluctuations of transient responses are greater or less than the same fluctuations for the evaluation  $\bar{J}_n^{(m)}_{\text{opt}}$ .

### 9. The Inversion of the Integral Evaluation $\bar{J}_n^{(m)}$ by means of the Spectra $R_n$ and $w_m$

To estimate the transient performance in a control system various integral criteria have found broad application. This is done most often by a comparative method. The less is the value of the integral evaluation chosen, the higher is the quality of the transient response. In this connection various methods have been developed for investigation of the relation between a change of values of selected transfer function parameters and the corresponding change of the value of the integral evaluation. Of the best known and most widely used, mention should be made of methods of minimizing the integral evaluation in relation to one or a few parameters: graphoanalytic methods of determining the minimum evaluation and the method of successive trials and approximations.

The aim is to obtain analytically a new solution of this problem using the integral square criterion of transient performance which was called the inverse problem of the integral square estimation  $\bar{J}_n^{(m)}$  and which could also be called the inverse problem of transient performance. This is a problem encountered particularly in the synthesis of linear systems.

Under the name of inversion of the integral square estimation  $\bar{J}_n^{(m)}$  one will understand the determination of the normalized transfer function for a prescribed value of the normalized Krasovskii integral evaluation  $\bar{J}_n^{(m)}$ . The solution of the inversion problem of the integral criterion  $\bar{J}_n^{(m)}$  has been obtained by introducing the notions of the spectra  $R_n$  and  $w_m$  defined above. It should be explained that in the general case the evaluation  $\bar{J}_n^{(m)}$  is a multivalued function. For any assigned value of the evaluation  $\bar{J}_n^{(m)}$  an infinite number of various linear transfer functions can be made to correspond. A different case is that where the evaluation  $\bar{J}_n^{(m)}$  is expressed in terms of the spectra  $R_n$  and  $w_m$ , the correspondence between a transfer function and these spectra being now one-to-one. In this connection, if the inverse evaluation  $\bar{J}_n^{(m)}$  is spoken of, one always means the inverse evaluation  $\bar{J}_n^{(m)}$  expressed in terms of definite spectra  $R_n$  and  $w_m$ .

Assume that the spectra of the Krasovskii evaluation  $\bar{J}_n^{(m)}$ , primary  $R_n = R_n(r_1, r_2, \dots, r_n)$  and secondary  $w_m = w_m(h_1, h_2, \dots, h_m)$ , are known. They contain full information on the

transfer function  $\bar{F}(p) = C(p)/A(p)$  of the control channel under consideration and much information on the transient performance in this control channel.

The transfer function will be determined by inversion of the evaluation  $\bar{J}_n^{(m)}$  expressed in terms of the spectra  $R_n$  and  $w_m$ ; that is, by inverse transformation of the spectra  $R_n$  and  $w_m$ . For this purpose determine first the characteristic equation corresponding to the spectrum  $R_n$ .

The method of determining the characteristic equation (polynomial)  $A(p) = p^n + a_1 p^{n-1} + \dots + a_{n-1} p + a_n$  is an elementary one. The values of all the elements  $r_1, r_2, \dots, r_n$  of the spectrum  $R_n$  being known, it is easy, on the basis of (29), to determine, for instance, the values of all the Hurwitz determinants and then to make use of Table 1 which enables the values of the coefficients of the characteristic polynomial  $A(p)$  to be found directly.

In order to avoid the intermediate stage of computing the Hurwitz determinants, Table 3 has been prepared, containing expansions of the coefficients of the characteristic equation directly in terms of the elements  $r_i$  ( $i = 1, 2, \dots, n$ ) of the primary spectrum  $R_n$ . In this case the coefficients  $a_{k,n}$  may be determined by means of the algorithm

$$a_{k,n} = a_{k,n-1} + \frac{a_{k-2,n-2}}{r_{n-1} \cdot r_n} \quad (53)$$

or by means of the recurrence equations

$$A_k(p) = p A_{k-1}(p) + \frac{A_{k-2}(p)}{r_{k-1} \cdot r_k} \quad (k = 1, 2, \dots, n) \quad (54)$$

The expressions (53) and (54) are equivalent to eqns (8) and (9)<sup>7</sup>. In the case of low degrees  $n$  in the characteristic polynomial it is more convenient to use Table 3. In the case of high degrees  $n$  eqns (53) or (54) are better suited for computation.

The method for finding the polynomial  $c(p)$  in the numerator of the transfer function  $\bar{F}(p) = C(p)/A(p)$  is also elementary; it consists in representing the polynomial  $C(p)$  as a sum of two polynomials

$$C(p) = c_m p^m + c_{m-1} p^{m-1} + \dots + c_1 p + 1 = C(p)_{\text{opt}} + \Delta C(p) \quad (55)$$

where

$$\left. \begin{aligned} C(p)_{\text{opt}} &= c_{m \text{ opt}} p^m + c_{m-1 \text{ opt}} p^{m-1} + \dots + c_{1 \text{ opt}} p + 1 \\ \Delta C(p) &= h_m p^m + h_{m-1} p^{m-1} + \dots + h_1 p \end{aligned} \right\} \quad (56)$$

The coefficients  $c_{\alpha \text{ opt}}$  ( $\alpha = 1, 2, \dots, m$ ) of the polynomial  $C(p)_{\text{opt}}$  can easily be obtained from Table 2 and the spectrum  $R_n$ . One has only to take into consideration the relation  $c_{\alpha \text{ opt}} = b_{m-\alpha \text{ opt}}$  between the coefficients of the equivalent polynomial  $C(p)_{\text{opt}}$  and  $B(p)_{\text{opt}}$ . The coefficients  $h_i$  ( $i = 1, 2, \dots, m$ ) are known if the spectrum  $w_m$  is known. As a consequence the transfer function  $\bar{F}(p) = C(p)/A(p)$  is uniquely determined. The spectra  $R_n$  and  $w_m$  enable one to estimate simultaneously the stability of a system and the transient performance in agreement with the principles studied in Sections 6–8.

The method based on the inverse integral criterion  $\bar{J}_n^{(m)}$  and the spectra of the evaluation  $\bar{J}_n^{(m)}$  may be used successfully for analysis and synthesis of linear systems of automatic control. In the first case the transfer function is known, therefore the

Table 3. Expansion of the coefficients of the characteristic equation in terms of the elements  $r_i$  ( $i = 1, 2, \dots, n$ ) of the primary spectrum  $R_n$  of the evaluation  $\bar{J}_n^{(m)}$ 

Coefficients of the characteristic equation $A(p) = a_0 p^n \div a_1 p^{n-1} \div \dots \dots a_n$								
$n$	$a_0$		$a_1$	$a_2$	$a_3$	$a_4$	$a_5$	$a_6$
	1	1	$\frac{1}{r_1}$					
2	1	1	$\frac{1}{r_1}$	$a_2 = \frac{1}{r_1 r_2}$				
3	1	1	$\frac{1}{r_1}$	$\frac{1}{r_1 r_2} + \frac{1}{r_2 r_3}$	$a_3 = \frac{1}{r^2} \left( \frac{1}{r_2 r_3} \right)$			
4	1	1	$\frac{1}{r_1}$	$\frac{1}{r_1 r_2} + \frac{1}{r_2 r_3} + \frac{1}{r_3 r_4}$	$\frac{1}{r_1} \left( \frac{1}{r_2 r_3} + \frac{1}{r_3 r_4} \right)$	$a_4 = \frac{1}{r_3 r_4} \left( \frac{1}{r_1 r_2} \right)$		
5	1	1	$\frac{1}{r_1}$	$\frac{1}{r_1 r_2} + \frac{1}{r_2 r_3} + \frac{1}{r_3 r_4} + \frac{1}{r_4 r_5}$	$\frac{1}{r_1} \left( \frac{1}{r_2 r_3} + \frac{1}{r_3 r_4} + \frac{1}{r_4 r_5} \right)$	$\frac{1}{r_3 r_4} \left( \frac{1}{r_1 r_2} \right) + \frac{1}{r_4 r_5} \left( \frac{1}{r_1 r_2} + \frac{1}{r_2 r_3} \right)$	$a_5 = \frac{1}{r_1 r_4 r_5} \left( \frac{1}{r_2 r_3} \right)$	
6	1	1	$\frac{1}{r_1}$	$\frac{1}{r_1 r_2} + \frac{1}{r_2 r_3} + \frac{1}{r_3 r_4} + \frac{1}{r_4 r_5} + \frac{1}{r_5 r_6}$	$\frac{1}{r_1} \left( \frac{1}{r_2 r_3} + \frac{1}{r_3 r_4} + \frac{1}{r_4 r_5} + \frac{1}{r_5 r_6} \right)$	$\frac{1}{r_3 r_4} \left( \frac{1}{r_1 r_2} \right) + \frac{1}{r_4 r_5} \left( \frac{1}{r_1 r_2} + \frac{1}{r_2 r_3} \right) + \frac{1}{r_5 r_6} \left( \frac{1}{r_1 r_2} + \frac{1}{r_2 r_3} + \frac{1}{r_3 r_4} \right)$	$a_5 = \frac{1}{r_1 r_4 r_5} \left( \frac{1}{r_2 r_3} \right) + \frac{1}{r_1 r_5 r_6} \left( \frac{1}{r_2 r_3} + \frac{1}{r_3 r_4} \right)$	$a_6 = \frac{1}{6} \prod_{i=1}^6 r_i$

Notes: (1)  $J_n^{(m)}$  is the Krasovskii integral evaluation.(2) In the case of normalized characteristic equation the condition  $\prod_{i=1}^n r_i = 1$  should be satisfied.

corresponding spectra of the evaluation  $\bar{J}_n^{(m)}$  can be found easily; in particular, the primary spectrum  $R_n$  can rapidly be determined by using, for instance, the Markov criterion or finding, by successive elimination, the elements  $r_i$  ( $i = 1, 2, \dots, n$ ) directly from the expansion of the characteristic equation in terms of these elements. The next stage is that of correction of the spectra obtained.

In the case of synthesis it is necessary to know the general structure of the transfer function (that is  $n$  and  $m$  must be known) and the requirements concerning the evaluation  $\bar{J}_n^{(m)} = \bar{J}_n^{(m)}_{\text{opt}} + \Delta M_n^{(m)}$ . Correct choice of the weight elements  $r_i$  ( $i = 1, 2, \dots, n - m$ ) and the elements  $h_i$  of the spectrum  $w_m$  is of particular importance. It follows that the problem of correct choice of the spectra  $R_n$  and  $w_m$  is essential for the synthesis of linear systems and the necessary correction of such system.

### 10. Minimization of the Integral Evaluation $\bar{J}_n^{(m)}$

Some additional data on the primary spectrum  $R_n$  may be obtained by minimizing the integral evaluation  $\bar{J}_n^{(m)}$ . The minimization of the evaluation  $\bar{J}_n^{(m)} = \bar{J}_n^{(m)}_{\text{opt}} + \Delta M_n^{(m)}$ , expressed in terms of the SMI may be divided into two stages. The first consists in obtaining the optimum integral estimation  $\bar{J}_n^{(m)}$  in the sense of SMI, discussed in Section 5. As result the component  $\Delta M_n^{(m)}$  vanishes.

The next stage consists in minimizing the component  $\bar{J}_n^{(m)}_{\text{opt}}$  in relation to a selected group of SMI, for instance

$$\begin{aligned} H &= H(\Delta_1, \Delta_2, \dots, \Delta_n) \\ R &= R(q_0, q_1, \dots, q_{n-1}) \\ M &= M(\bar{q}_0, \bar{q}_1, \dots, \bar{q}_{n-1}) \end{aligned} \quad (57)$$

where  $H, R, M$  are groups of determinant indices of stability margin in the sense of the criterion of Hurwitz, Routh and Markov, respectively. Of course, the result of the minimization process depends on the choice of the group of independent variables  $H, R$  or  $M$ .

As a result of investigations a new property of the spectrum has been found, characteristic of the conditional minimum of the evaluation  $\bar{J}_n^{(m)}$ . This is as follows:

**Property III:** An integral evaluation  $\bar{J}_n^{(m)}$  minimized with respect to an appropriate group of SMI reaches a conditional minimum if its secondary spectrum  $w_m$  is zero and the weight elements of the primary spectrum  $R_n$  constitute an arithmetic progression of which the difference is  $\Delta r_i$  and  $\Delta r_i = \kappa/(n - m)\Delta_1$ ;  $r_{i+1} = r_i + \Delta r_i$  ( $i = 1, 2, \dots, n - m$ ;  $m > 0$ ).

The conditional minimum of the evaluation  $\bar{J}_n^{(m)}$  ( $m > 0$ ) is expressed by the equation (Jarominek<sup>7</sup>),

$$\bar{J}_n^{(m)}(\kappa)_{\min} = \frac{2(n - m) + \kappa(n - m - 1)}{4\Delta_1} > 0 \quad (58)$$

In the particular case of

$$\kappa = -1 \text{ one has } \bar{J}_n^{(m)}(H)_{\min} = \bar{J}_n^{(m)}(\kappa = -1)_{\min} = \frac{n - m + 1}{4\Delta_1} \quad (59)$$

and

$$\kappa = 0 \text{ one has } \bar{J}_n^{(m)}(R)_{\min} = \bar{J}_n^{(m)}(\kappa = 0)_{\min} = \frac{n - m}{2\Delta_1} \quad (60)$$

### 11. The Problem of Correct Distribution of Zeros of the Transfer Function

The secondary spectrum  $w_m$  characterizes indirectly the distribution of zeros of the transfer function, which has a strong influence on the value of the evaluation  $\bar{J}_n^{(m)}$ . As a criterion of correct distribution of the zeros of the transfer function in relation to the distribution of the poles, the value of the component  $\Delta M_n^{(m)} \geq 0$  of the evaluation  $\bar{J}_n^{(m)} = \bar{J}_n^{(m)}_{\text{opt}} + \Delta M_n^{(m)}$  is taken. The component  $\Delta M_n^{(m)}$  depends on both spectra  $R_n$  and  $w_m$ . The spectrum  $w_m$ , which is non-zero in general, will be correctly chosen to fit the spectrum  $R_n$  if the component  $\Delta M_n^{(m)}$  does not exceed a certain fixed value.

The upper bound of the value of the component  $\Delta M_n^{(m)}$  can be determined assuming as a principle the maximum utilization of the polynomial  $C(p)$  of degree  $m$ . For this will be used the first equivalent form of the integral criterion  $\bar{J}_n^{(m)}$ , that is

$$\bar{J}_n^{(m)} = \bar{J}_n^{(0)} + M_n^{(m)} \text{ and } \bar{J}_n^{(m)}_{\text{opt}} = \bar{J}_n^{(0)} + M_n^{(m)}_{\text{opt}} \quad (61)$$

On the basis of appropriate considerations the following can be written:

$$\liminf M_n^{(m)} = M_n^{(m)}_{\text{opt}} \leq M_n^{(m)} \leq \limsup M_n^{(m)} = M_n^{(m-1)}_{\text{opt}} \quad (62)$$

It follows that

$$0 \leq M_n^{(m)} - M_n^{(m)}_{\text{opt}} = \Delta M_n^{(m)} \leq M_n^{(m-1)} - M_n^{(m)}_{\text{opt}} \quad (63)$$

On the other hand it is easy to show that

$$M_n^{(m-1)} - M_n^{(m)}_{\text{opt}} = \frac{1}{2} r_{n-m+1} \quad (64)$$

It is inferred that  $\Delta M_n^{(m)}$  should be contained between the limits

$$0 \leq \Delta M_n^{(m)} \leq \frac{1}{2} r_{n-m+1} \quad (65)$$

The secondary spectrum  $w_m$  is therefore correctly chosen if condition (65) is satisfied.

The above method, based on the solution of the inverse stability problem and the inverse transformation of the integral square estimation  $\bar{J}_n^{(m)}$  may also be successfully used for the analysis and synthesis of multi-loop linear systems of automatic control. In the case of synthesis it enables the parameters following the requirements concerning the static and dynamic characteristics of the system to be chosen<sup>7</sup>. It may also be helpful for the investigation of non-linear systems in the applicability limits of the Liapunov theorems<sup>10</sup>, and for the investigation of some adaptive systems.

### References

- JAROMINEK, W. Investigating linear systems of automatic control by means of determinant stability margin indices. *Automatic and Remote Control*. 1960. London; Butterworths
- KRASOVSKII, A. A. *Integral Evaluations and the Selection of Parameters of Automatic Control Systems*. 1954. Moscow; Mashgiz
- POPOV, E. P. *Dinamika Sistem Avtomaticheskovo Regulirovania*. 1954. Gostekhizdat
- JAROMINEK, W. Ob ekvivalentnosti kriteria ustoichivosti po Routh'u, Hurwitz'u i po Markov'u. *DAN U.S.S.R.*, T. 130, No. 6 (1960)

- <sup>5</sup> JAROMINEK, W. Primenenie pokazatelei zapasa ustoichivosti po opredelitelam dla issledovaniia lineinykh sistem avtomaticheskogo regulirovaniia (1959)
- <sup>6</sup> JAROMINEK, W. Razlozhenie koeffitsientov kharakteristicheskogo uravnenia po determinantnym pokazateliam zapasa ustoichivosti (in press)
- <sup>7</sup> JAROMINEK, W. Issledovanie lineinykh stationarnykh sistem metodom spektrov integralnykh ocenok (1961)
- <sup>8</sup> JAROMINEK, W. Ob odnom sluchae nakhozhdeniia minimuma kvadraticnoi integralnoi ocenki. *Izv. Vyzshikh Uchebnykh Zavedenii, razdel Electromech.* No. 13 (1959)

- <sup>9</sup> JAROMINEK, W. O zapase ustoichivosti lineinykh sistem avtomaticheskogo regulirovaniia. *Izv. Vyzshikh Uchebnykh Zavedenii, razdel Electromech.*, No. 8 (1959)
- <sup>10</sup> LETOV, A. M. Sostoianie problemy ustoichivosti v teorii avtomaticheskogo regulirovaniia, *T. I. Izd. AN U.S.S.R.* (1955)
- <sup>11</sup> LEJA, F. Rachunek rozniczkowy, calkowy, *PWN*, Warsaw (1954)
- <sup>12</sup> PETROV, B. N. *Sviaz Mezhdru Kachestvom Perekhodnogo Processa i Pazpredeleniem nulei i Poliusov Peredachnoi Funkcii v sb; Teoria Avtomaticheskogo Regulirovaniia.* 1954. Moscow; Mashgiz

## DISCUSSION

P. C. PARKS, *Department of Aeronautics, The University, Southampton, England*

There are some interesting connections between the work of Jarominek in this paper and in his earlier work (Reference 1 of the paper) with Liapunov's second method, and work by Schwarz<sup>1</sup> in Switzerland and Kalman and Bertram<sup>2</sup> in the U.S.A. If one considers the possibility of writing eqn (3) in the special state space form

$$\dot{x} = Bx$$

where  $B$  is the Schwarz matrix

$$B = \begin{matrix} & 0 & 1 & 0 & 0 \dots & 0 & 0 & 0 \\ -b_n & 0 & 1 & 0 \dots & 0 & 0 & 0 & 0 \\ 0 & -b_{n-1} & 0 & 1 \dots & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \dots & -b_3 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 \dots & 0 & -b_2 & -b_1 & 0 \end{matrix}$$

then Parks<sup>3</sup> has shown that

$$b_1 = \Delta_1, \quad b_2 = \frac{\Delta_2}{\Delta_1}, \quad b_3 = \frac{\Delta_3}{\Delta_1 \Delta_2}, \quad b_r = \frac{\Delta_r - \Delta_{r-3} \Delta_r}{\Delta_{r-2} \Delta_{r-1}} \quad (r \geq 4)$$

where the  $\Delta_r$  are the Hurwitz determinants of (3). For the state space equation above, Kalman and Bertram<sup>2</sup> found the Liapunov function

$$V = b_1 b_2 \dots b_n x_1^2 + b_1 b_2 \dots b_{n-1} x_2^2 + \dots + b_1 x_n^2$$

for which  $\dot{V} = -2 b_1^2 x_n^2$  which is negative-semidefinite and not identically zero along any trajectory, except the origin itself. Necessary and sufficient stability conditions follow from the positiveness of the  $b_i$  and hence, from the work of Parks<sup>3</sup>, the  $\Delta_i$ . Moreover, the expansion of  $\det(\lambda I - B)$  from the bottom right-hand corner leads to a solution of the 'inverse stability problem', using the expressions for the  $b_i$  in terms of the  $\Delta_i$ .

Further, if one considers

$$\dot{x} = Bx + df(t)$$

where  $d = (0 \ 0 \ 0 \dots 0 \ 1)$  and  $f(t)$  is a unit step function then

$$\bar{J}_n^{(0)} \int_0^\infty x_n^2 dt = \frac{1}{2 b_1^2} \{V(0) - V(\infty)\} = \frac{1}{2 b_1^2} V(0) \text{ as } V(\infty) = 0$$

for a stable system. The state space motion decays to the origin from the starting point

$$x_n = 1, \quad x_{n-1} = -\frac{b_1}{b_2}, \quad x_{n-2} = \frac{1}{b_3}, \quad x_{n-3} = -\frac{b_1}{b_2 b_4},$$

$$x_{n-4} = \frac{1}{b_3 b_5}, \quad x_{n-5} = -\frac{b_1}{b_2 b_4 b_6} \text{ etc.}$$

and

$$\bar{J}_n^{(0)} = \int_0^\infty x_n^2 dt = \frac{1}{2 b_1^2} \left\{ b_1 + b_1 b_2 \frac{b_1^2}{b_2^2} + b_1 b_2 b_3 \frac{1}{b_3^2} + \dots \right\}$$

$$= \frac{1}{2} \left\{ \frac{1}{\Delta_1} + \frac{\Delta_1^2}{\Delta_2} + \frac{\Delta_2^2}{\Delta_1 \Delta_3} + \frac{\Delta_3^2}{\Delta_2 \Delta_4} + \dots + \frac{\Delta_{n-1}^2}{\Delta_{n-2} \Delta_n} \right\}$$

which is Jarominek's expression (14). This is minimized for the normalized system

$$\left( a_{n,n} = \frac{\Delta_n}{\Delta_{n-1}} = 1 \right)$$

when  $\Delta_i = +1$  ( $i = 1, 2, \dots, n-1$ ) as mentioned by Jarominek in his earlier paper. This leads to 'optimum' transfer functions<sup>4</sup> and corrections to the analogue computer work of Graham and Lathrop<sup>5</sup>. There are further interesting ramifications of the Liapunov function  $V$  when a linear transformation to the phase space is carried out<sup>6, 7</sup>.

## References

- <sup>1</sup> SCHWARZ, H. R. *Z. angewandte Math. Phys.* 7 (1956), 473-500
- <sup>2</sup> KALMAN, R. E. and BERTRAM, J. E. *Trans. A.S.M.E. (Series D)* 82 (1960), 371-393
- <sup>3</sup> PARKS, P. C. *Proc. Cambridge Phil. Soc.* 58 (1962) 694-702
- <sup>4</sup> PARKS, P. C. *Proc. Joint Autom. Contr. Conf.* Minneapolis (1963) 471-478
- <sup>5</sup> GRAHAM, D. and LATHROP, R. C. *Trans. A.I.E.E.* 72, Pt. 2 (1953) 278-288
- <sup>6</sup> PARKS, P. C. *Trans. I.E.E.E. on Automatic Control* AC-8 (1963) No. 3 (Correspondence)
- <sup>7</sup> PARKS, P. C. Discussion on paper by Merklinger, *Proc. 2nd I.F.A.C. Congr.* Basle 1963. London; Butterworths: Munich; Oldenbourg

W. JAROMINEK, *in reply*

The results presented by Mr. Parks are very interesting. Especially valuable is the indication that there exists a connection between some of my results and works of Schwarz (Switzerland) and Kalman and Bertram (U.S.A.). Mr. Parks has shown that using Liapunov's second method it is possible, on the grounds of works of the above-mentioned authors, to confirm the correctness of some of the expressions obtained in a different way in my earlier works.

From the methodological point of view, I consider as most valuable the application by Mr. Parks of the second method of Liapunov to solutions of certain problems dealt with in my works, which opens new and interesting prospects of further development of the considered theory of automatic control.

# On Systems with Automatic Control of Configuration

J. BENEŠ

## Summary

The definition, the scheme and five basic operations of systems with automatic control of configuration are given. For the description of a complex with many elements, division into zones and statistical characteristics as state variables are introduced. The configurational redundancy and simpler measurable quantities are order measures. The probabilities involved in the configuration are influenced by the formator through controlled probabilistic transducers. The methodical approach to the solution of the formator and complex interaction consists in the formation of a mathematical model, checked against physical measurements and in using it in the choice of the control algorithm. The two basic deviations and cubic matrices of variables of this type of multivariable systems are introduced. The optimization of the process of configuration is described using the method of Howard in terms of a Markov process with decisions and rewards. Three examples of systems are suggested, pertaining to formation of strips of elements and to the migration of elements in one-dimensional and two-dimensional arrays of zones. Further progress is expected from physical modelling of non-homogeneous Markov processes. The field of application of the theory can be seen in chemistry, in automatic cultivation of algae and in random net formation.

## Sommaire

On donne la définition, le schéma et cinq opérations de base des systèmes avec la commande automatique de la configuration. Pour la description du complexe avec un grand nombre d'éléments on introduit sa division en zones et des caractéristiques statistiques comme variables décrivant son état. La redondance de configuration et des quantités mesurables plus simples sont des mesures de l'ordre. Les probabilités concernant la configuration sont influencés par le formateur par l'intermédiaire de convertisseur commandes de probabilité. L'accès méthodique à la solution de l'interaction du formateur et du complexe réside dans la formation d'un modèle mathématique qui est à vérifier par des mesures physiques, et dans son emploi pour le choix de l'algorithme de commande. On introduit les deux écarts de base et les matrices cubiques des variables de ce type de systèmes à variables multiples. On décrit l'optimisation du processus de configuration, en employant la méthode de Howard, en tant que processus de Markov avec des décisions et des récompenses. On propose trois exemples, concernant la formation de chaînettes d'éléments et la migration des éléments dans des dispositifs unidimensionnels et bidimensionnels de zones. On attend un nouveau progrès de la création de modèles physiques de processus non-homogènes de Markov. Le champ d'application de la théorie peut être entrevu en chimie, pour la culture automatique des algues et pour la formation de réseaux aléatoires.

## Zusammenfassung

Die Definition, das Schema und fünf Grundarbeitsaufgaben von Systemen mit selbsttätiger Regelung der Konfiguration werden gegeben. Für die Beschreibung des Komplexes mit vielen Elementen wird seine Einteilung in Zonen und statistische Charakteristiken als Zustandsgrößen eingeführt. Die Konfigurationsredundanz und einfachere Meßgrößen sind Maße der Ordnung. Die mit der Konfiguration verknüpften Wahrscheinlichkeiten werden vom Formator durch gesteuerte Wahrscheinlichkeitswandler beeinflusst. Der methodische

Zutritt zur Lösung der gegenseitigen Tätigkeit des Formators und des Komplexes liegt in der Bildung eines mathematischen Modells, das durch physikalische Messungen zu prüfen ist, und in seiner Anwendung zur Wahl des Algorithmus für die Regelung. Zwei Grundabweichungen und kubische Matrizen von Variablen von diesem Typus von Systemen mit vielen Veränderungsgrößen werden eingeführt. Die Optimierung des Konfigurationsprozesses wird beschrieben, mit Anwendung der Methode von Howard, als eines Prozesses von Markoff mit Entscheidungen und Belohnungen. Drei Beispiele dieser Systeme werden vorgeschlagen, betreffend die Bildung von Ketten von Elementen und die Wanderung von Elementen in ein- und zwei-dimensionalen Zonenanordnungen. Weiterer Fortschritt wird von dem physikalischen Modellieren von unhomogenen Markoffschen Prozessen erwartet. Das Anwendungsgebiet der Theorie kann in der Chemie, beim selbsttätigen Kultivieren von Algen und in der Bildung von zufallsbestimmten Netzen gesehen werden.

## Introduction

A further development of the theory of automatic control results from its application to complexes of many elements subject to automatic ordering action. Owing to the impossibility of following the dynamics of the very numerous elements of the complex, statistical characteristics, accessible to macroscopic measurement, may be used for the description of the evolution of the group of elements. These characteristics are to be compared with the corresponding theoretical ones, derived from the mathematical model of the process of configuration based upon the properties of the elements and upon the conditions of the process, including the influence of the control action. The problem of controlling the development of the group of elements leads to the introduction of a deterministic control of certain frequency functions of events pertaining to the formation of configurations. The corresponding mathematical models use probabilities instead of frequency functions.

## Definition and Basic Scheme

A system with automatic control of configuration is a system with a complex of elements which develops by automatic control towards an assigned state or set of states, characterized by the configuration of these elements.

During this development one (or more) of the following basic operations of configuration, pertaining to the elements of the complex, is realized: the aggregation; the orientation; the liaison; the arrangement; the connection.

The general scheme of a system with automatic control of configuration is shown in Figure 1. Here  $K$  = the complex;  $F$  = the formator;  $S$  = the measured state variables of the complex;  $A$  = the acting variables of the formator;  $P$  = the perturbing signals acting upon the complex;  $V$  = the output variables of the complex;  $R$  = the command variables.



The function of the formator is to elaborate the acting signals for influencing the configuration of the elements of the complex. The output variables  $V$  of the complex are, in general, different from the measured state variables  $S$  which are chosen so as to inform, by their totality, on the configuration of the elements of the complex.

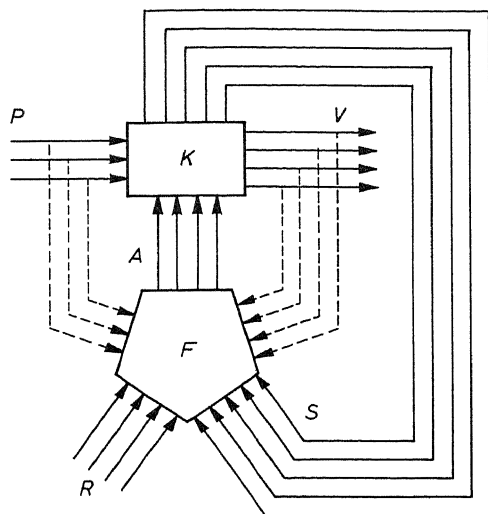


Figure 1

### Description of the State of the Complex

An approach to the description of a complex with a great number of elements consists in its *division into equal zones* and in considering the group of elements contained in each of these zones. The interrelation of the groups contained in the zones, especially in the neighbouring ones, may be of interest. In two-dimensional representation we draw the meshwork of zones as, for example, in Figures 7 and 8. As it may not be possible to measure the state variable in all these zones, we introduce *sample zones*. The measured state of the complex may be expressed by the measured state variables  $S$  in the form: (a) of a column vector with elements  $s_1(t), s_2(t), \dots, s_n(t)$ ; (b) or, in the case of a two-dimensional arrangement of measuring points, which may be advantageous for the expression of the configuration, in the form of a quadratic matrix  $\|s_{ij}(t)\|$  ( $i, j = 1, 2, \dots, n$ ); (c) or, in the case of a three-dimensional arrangement of measuring points, in the form of a cubic matrix

$$\|s_{ijk}(t)\| \quad (i, j, k = 1, 2, \dots, n) \quad (1)$$

Measurements of state variables in three-dimensional or in two-dimensional arrays of zones of the complex can be reduced by scanning to a sequence of measurements. Similarly, one can express the required state of the complex using the command variables  $R$ .

A theoretical measure of the state of the complex with many elements is the configurational redundancy

$$R_{fm} = 1 - \frac{\Delta S_{fm}}{\Delta S_{fv}} \quad (2)$$

where  $\Delta S_{fm} < \Delta S_{fv}$ .

The index  $m$  applies to the intermediate state between the initial state (index  $v$ ) and the final state (index  $c$ ) and where the differences of configurational entropy are

$$\Delta S_{fm} = k \log(Z_m - Z_c) \quad (3)$$

$$\Delta S_{fv} = k \log(Z_v - Z_c) \quad (4)$$

where  $k$  is a scale factor.  $Z_v$  is the number of possible ways of having the elements ordered at the initial state,  $Z_m$  is the number of different ways of ordering the elements suiting the definition of the intermediate state at a certain phase of development and  $Z_c$  is the number of different ways of ordering the elements suiting the requirements upon the final state.

The configurational entropy is a concept used in statistical physics. In crystallography, the entropy change for a transition in the crystalline phase is divided into: the change of the configurational entropy and the change of thermal entropy<sup>8</sup>. The configurational entropy of the arrangement of atoms in a lattice is determined by the number of different ways in which the atoms may be arranged over the available number of lattice sites. In chemistry, the information content  $I_t$  of a protein is divided into:  $I_s$  depending upon the amino acid sequence, and  $I_c$  depending upon the configuration of the polypeptide chain in the native molecule<sup>9</sup>.

The state of the complex may be described by different concepts and measures depending upon the basic operation of configuration of the elements<sup>6, 12</sup>.

### Quantitative Expression of State During Aggregation

The simplest expression of the state is in terms of the number of elements or of their concentration in the different zones. The information connected with the concentration into a single zone  $s$  of elements of a certain type  $i$ , which previously have been distributed over the whole complex, is

$$I_{ci} = \log_2 \left( \frac{c_{is}}{c_{ik}} \right) \quad [\text{bit}] \quad (5)$$

where  $c_{is}$  is the concentration of the elements of type  $i$  in the zone  $s$ ,

$c_{ik}$  is the concentration of the elements of type  $i$  in the whole complex.

When several types  $i$  ( $i = 1, 2, \dots$ ) of elements are involved, we have

$$I_c = \sum_i \log_2 \left( \frac{c_{is}}{c_{ik}} \right) \quad [\text{bit}] \quad (6)$$

### Quantitative Expression of State During Orientation

The orientation of the elements of a zone of the complex may be expressed by an angular measure. The information connected with the orientation of the  $i$ th element of the complex may be expressed by its orientation information  $I_{or[i]}$ . If the configuration of the complex requires that the orientation of the  $i$ th element be fixed within  $\Delta\Theta, \Delta\phi, \Delta\psi$ , where  $\Theta, \phi, \psi$  are Euler angles, we have

$$I_{or[i]} = \log_2 \left\{ \frac{8\pi^3}{\Delta\Theta_i \Delta\phi_i \Delta\psi_i} \right\} \quad [\text{bit}] \quad (7)$$

### Quantitative Expression of State During Liaison

Consider the operation of liaison of the elements in a complex of constant volume, with elements of different types  $i$ , where  $i = 1, 2, \dots, k$  and denote  $n_1, n_2, \dots, n_k$  the numbers of these elements. They move and combine at random to form new types of elements by liaison. To characterize the development of the state of the complex use the probability  $P(\mathbf{n}, t)$ , which is the probability that in time  $t$  the complex has the composition  $\mathbf{n}$ , where  $\mathbf{n}$  is a vector, whose components are the numbers of the elements of the different types. By the action of the formator one wishes to influence the probability  $P(\mathbf{n}, t)$ .

### Quantitative Expression of State During Arrangement

The number of sites or lattice sites correctly occupied by the elements of the complex is a simple measure of arrangement. The information connected with the position of the  $i$ th element in the complex of volume  $V$  may be expressed by its placement information  $I_p[i]$ . If the configuration of the complex requires that the  $i$ th element remain within a space  $\Delta x_i, \Delta y_i, \Delta z_i$ , we have

$$I_p[i] = \log_2 \left\{ \frac{V}{\Delta x_i \Delta y_i \Delta z_i} \right\} \quad [\text{bit}] \quad (8)$$

### Quantitative Expression of State During Connection

As a characteristic quantity of a random net, Clark and Farley have used the connectivity. An element,  $i$ , is connected to an element  $j$  with a probability  $P_{ij}$  which may depend upon both  $i$  and  $j$  and on other characteristic quantities of the net as a whole. Uttley has considered the probability of the connection of an element in a given position to an input point. This probability is a function of the position. Beurle has used a probability of connection of an element to all elements which are in a distance  $r$  from it, this probability being a function of the co-ordinates  $x, y, z$  of the element and of the distance  $r$ .

As we are interested in characterizing the state of a developing random net, the use of test impulses, applied in points of sample zones and the measurement of the number of elements activated at a given distance in a given direction, or of the speed of the signal spreading, can be suggested, according to the particular case. These quantities are related to the statistical characteristics of the random net.

### The Interaction of the Formator and of the Complex

The acting variables of the formator are from the point of view of automatic control the output variables of a multi-dimensional controller with many inputs, some of which are the measured state variables of the complex. The information about the behaviour of the complex only from the measurement of external input and output variables of the complex would be insufficient. This is also in compliance with the principle of uncertainty in the structural behaviour of multivariable systems, formulated by Mesarović<sup>1</sup>.

Therefore direct measurement of the state variables, amended by theoretical relations yielded from the mathematical model of the configuration process, is required.

A methodical approach towards the identification of the process occurring in the complex consists in the following stages:

(1) The formation of a mathematical model of the process of configuration of the elements.

(2) Computation of the relevant mean values  $S^{(T)}(t)$  on a mathematical machine on the basis of the mathematical model and using complementary information about the physical conditions of the process.

(3) Comparison of the computed mean values  $S^{(T)}$  and their development in time with the measured state variables  $S$ .

(4) The appropriate adaptive correction of the model under 2 in order to minimize the difference of the comparison under 3.

The objective is to obtain an approximate model of the process in the complex based on theoretical results from the statistical dynamics of the elements of the complex. Such a model is intended to bring a better insight into the mechanism of configurational changes and to help in choosing the method of control.

Two basic deviations which have to be considered in the theory of these systems are:

$$(1) \quad \varphi(t) = R(t) - S(t) \quad (9)$$

where  $R(t)$  are the command variables and  $S(t)$  the measured state variables,

$$(2) \quad \varphi^{(T)}(t) = S^{(T)}(t) - S(t) \quad (10)$$

where  $S^{(T)}(t)$  are theoretical state variables computed from the mathematical model. The minimization of  $\varphi^{(T)}(t)$  is a form of the identification problem of the process in the complex.

When using the cubic matrix expressions the basic deviations are:

$$(a) \quad \|\phi_{ijk}(t)\| = \|r_{ijk}(t)\| - \|s_{ijk}(t)\| \quad (11)$$

$$(b) \quad \|\phi_{ijk}^{(T)}(t)\| = \|s_{ijk}^{(T)}(t)\| - \|s_{ijk}(t)\| \quad (i, j, k = 1, 2, \dots, n) \quad (12)$$

Consider a set of discrete simultaneous values of the time functions. A typical operation with the cubic matrix consists in applying to the trilinear form

$$F = \sum_{i, j, k=1}^n \phi_{ijk} x_i y_j z_k \quad (13)$$

the linear transformation

$$x_\gamma = \sum_{i=1}^n g_{\gamma i} X_i \quad (\gamma = 1, 2, \dots, n) \quad (14)$$

with the quadratic matrix

$$g = \|g_{\gamma i}\| \quad (\gamma, i = 1, 2, \dots, n) \quad (15)$$

There is then obtained the product cubic matrix in the index  $i$ :

$$\phi\{i\} g = \left\| \sum_{\gamma=1}^n \phi_{\gamma j k} g_{\gamma i} \right\| \quad (16)$$

Similarly,

$$\phi\{j\} g = \left\| \sum_{\gamma=1}^n \phi_{i \gamma k} g_{\gamma j} \right\| \quad (17)$$

$$\phi\{k\} g = \left\| \sum_{\gamma=1}^n \phi_{i j \gamma} g_{\gamma k} \right\| \quad (18)$$

Applying certain bilinear transformations with a cubic matrix of  $n$ th order to the trilinear form we would get product tetric matrices of the  $n$ th order, which are already more complex to handle<sup>3</sup>.

Owing to the number of input variables of the formator and to the complexity of its function, the formator should be a digital computer.

A special role in the introduction of this point of view of influencing probabilities of events by the formator is to be assigned to the concept of controlled probabilistic transducers.

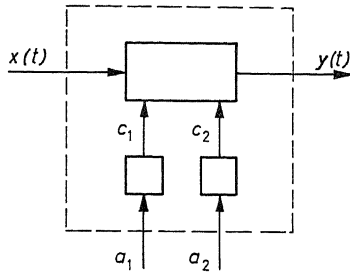


Figure 2

As an example of a type of controlled probabilistic transducer let us derive from the transducer type, mentioned by Köchen<sup>11</sup>, the transducer of Figure 2, where the conditional probabilities  $c_1$  and  $c_2$  are, respectively, functions of the acting variables  $a_1, a_2$  of the formator. The conditional probabilities are

$$c_1 = P[y(t) = 1 | x(t-1) = 1] \quad (19)$$

$$c_2 = P[y(t) = 1 | x(t-1) = 0] \quad (20)$$

where  $x(t)$  is a two-valued variable (value 1 or zero), and  $t$  denotes the number of the time interval.

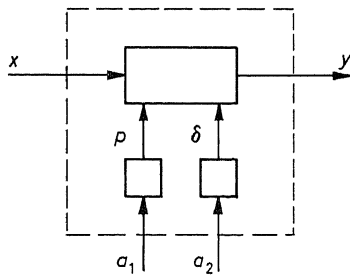


Figure 3

Another type of controlled probabilistic transducer (Figure 3) can be derived from the concept of binomial probabilistic transform, studied by Sugimori<sup>4</sup>, by making the parameters  $p$  and  $\delta$  functions of the acting variables  $a_1$  and  $a_2$  of the formator. Let  $x, y$  be the numbers of signals of value 1 and  $\delta$  a time lag. For the special case, when  $p$  and  $\delta$  are constants, and when  $x$  has a Poisson distribution with parameter  $\lambda$

$$p(x) = \frac{\lambda^x}{x!} e^{-\lambda} \quad (21)$$

then in time  $\delta$  and only in this time

$$p(y) = \sum_{x=0}^{\infty} p(x) b(y; x, p) = \frac{(\lambda p)^y}{y!} e^{-\lambda p} \quad (22)$$

where  $b(y; x, p)$  is the binomial distribution of the random variable  $S_x$  denoting the number of successes in  $x$  Bernoulli trials.

Further, a simple scheme of a controlled probabilistic transducer without time-lag for a sequence of equidistant pulses can be represented by a logical product element, steered by a generator of random pulses with controlled statistical parameter, under the assumption of the synchronization of the pulse sequences.

### The Optimization of the Process of Configuration

The attainment of a certain state of configuration (or set of states) can be considered as a result of two opposite actions: the one of configurational ordering, the other of disordering.

An important problem in connection with the systems with automatic control of configuration can be raised: the stability of the controlled complex in remaining in a definite set of states. For the description of the development of some multivariable complexes, mathematical models based on Markov processes seem suitable.

The number of states which the multivariable complex can take is infinite and countable. The theory of Markov processes with an infinite and countable number of states is to be applied.

In order to formulate certain basic relations concerning the optimization of the process of automatic configuration we shall purposely limit ourselves to consider it as a Markov process with a finite number  $N$  of states. Let this process be defined by the matrix of transition probabilities  $P = [p_{jk}]$  and by the matrix of rewards  $W = [w_{jk}]$ , where the indexes  $j$  and  $k$  apply to the transition from the state  $j$  to the state  $k$ .

The result  $w_{jk}$  is the increment of configurational ordering associated with this transition. It is to be expressed in units of configurational measure. A physical interpretation may be, for example, the increase in the number of a new type of element formed by the liaison of two types of elements, or of the number of correctly occupied sites in a lattice etc. During the development of the complex some  $w_{jk}$  can be negative.

Applying the method of Howard<sup>2</sup>, one expresses the mean reward from a transition

$$g = \sum_{j=1}^N \pi_j q_j \quad (23)$$

where  $\pi_j$  = the probability of the complex being in state  $j$  after a large number of steps,

$q_j$  = the immediate reward expected at state  $j$ , i.e. the expected reward connected with the transition of the complex from the state  $j$  to the next one.

The immediate result  $q_j$  is

$$q_j = \sum_{k=1}^N p_{jk} w_{jk} \quad (j=1, 2, \dots, N) \quad (24)$$

The aim is to maximize the mean reward  $g$  of the Markov process of configuration.

Supposing that at each state  $j$  one of the alternatives of the action of the formator upon the complex can be chosen. To each alternative  $a$  corresponds a transition probability  $p_{jk}^a$  and a reward  $w_{jk}^a$ . When chosen, the alternative becomes a decision  $d$ . One denotes by  $d_j(n)$  the decision taken at the state  $j$ , which

means  $n$  states before the attainment of the final state. A column vector  $\mathbf{d}$ , with elements  $d_j(n)$  expresses the chosen policy. The total expected reward during the development of the complex in  $n$  steps starting from the state  $j$  and applying a specific policy is  $v_j(n)$ . Under the assumption of the Markov process being completely ergodic, we have for large  $n$

$$v_j(n) = ng + v_j \quad (j=1, 2, \dots, N) \quad (25)$$

Between the introduced quantities there is the relation

$$g + v_j = q_j + \sum_{k=1}^N p_{jk} v_k \quad (j=1, 2, \dots, N) \quad (26)$$

Following further the method of Howard, let  $v_N = 0$  and call  $v_j$  the relative values of the policy. By a judicious choice of the  $p_{jk}$  and  $q_j$  for each state  $j$  the reward  $g$  is to be maximized.

(1) For each state  $j$  the alternative  $a'$  which maximizes the value [see the relation (24)]

$$q_j^a + \sum_{k=1}^N p_{jk}^a v_k \quad (27)$$

is to be determined. Here the index  $a$  denotes the values belonging to the alternative  $a$ . Then by putting  $p_{jk}^{a'} = p_{jk}$ ;  $q_j^{a'} = q_j$ , the resulting values are used below.

(2) Using these  $p_{jk}$  and  $q_j$  in the system of linear simultaneous equations (26) and by solving this system one gets the  $v_j$  and the  $g$  which will be again used in (1). By the iterative computation process involving (1) and (2) one finally gets the  $g$ , and the  $p_{jk}$  and  $q_j$ . The required speed of computation would be high. On the other hand, assuming that the complex evolves relatively slowly, the change of the  $p_{jk}$  would be done by the action variables of the formator. Without the action of the formator, the isolated complex would develop 'spontaneously' with transition probabilities  $p_{jk}^{(s)}$ .

Theoretical investigations require the application of the theory of Markov processes with an infinite and countable number of states. Some notions are common with the theory of Markov processes with a finite number of states, as for example, the notion of undecomposable groups, of transition groups and of final groups.

### Suggested Examples of Systems

As an example of a system with automatic control of configuration, consider a system with controlled operation of liaison of three different types  $A, B, C$  of the very numerous elements of its complex. The elements move with random Brownian movement and have the following properties: (i) when  $A$  and  $B$  collide, a new element  $D$  results, (ii) when  $D$  and  $C$  collide, a new element  $E$  results, (iii) the collisions of the elements are at random, (iv) the liaisons are irreversible. (v) the direct liaison of

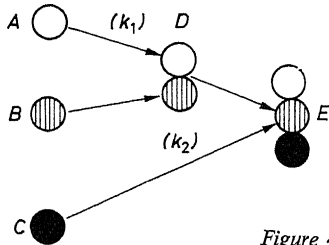


Figure 4

$C$  with  $A$  or  $B$  is impossible, (vi) the liaison rate parameters are  $k_1$  and  $k_2$  (Figure 4).

The state variables of the complex at time  $t = 0$  are  $n_{a0}, n_{b0}, n_{c0}, 0, 0$ —the numbers of elements of the different types. The command variables  $r_1, r_2$  are the required numbers of the elements  $D$  and  $E$  respectively (Figure 5). The acting variables of the formator are  $a_1, a_2$ , influencing the liaison rate parameters  $k_1$  and  $k_2$  respectively. Let  $x_1, x_2$  be the number of elements  $D$  and  $E$  respectively at time  $t$ .

The control of the operation of liaison is based on making the liaison rate parameters appropriate functions of time:  $k_1(t), k_2(t)$ .

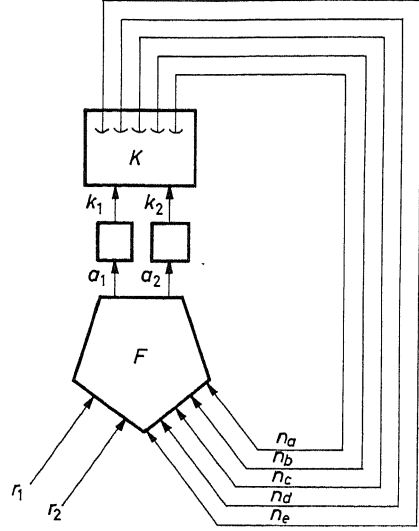


Figure 5

In order to simplify the expressions, first consider  $k_1$  and  $k_2$  as constants in the mathematical model. The differential equation describing the development of the complex is then:

$$\begin{aligned} \frac{dP(x_1, x_2, t)}{dt} = & k_1 (n_{a0} - x_1 + 1) (n_{b0} - x_1 + 1) P(x_1 - 1, x_2, t) \\ & - k_1 (n_{a0} - x_1) (n_{b0} - x_1) P(x_1, x_2, t) \\ & + k_2 (x_1 + 1) (n_{c0} - x_2 + 1) P(x_1 + 1, x_2 - 1, t) \\ & - k_2 (x_1) (n_{c0} - x_2) P(x_1, x_2, t) \end{aligned} \quad (28)$$

Using the method of the generating function one takes

$$F(s_1, s_2; t) = \sum_{x_1, x_2=0}^{\infty} P(x_1, x_2, t) s_1^{x_1} s_2^{x_2} \quad (29)$$

and finds

$$\begin{aligned} \frac{\partial F}{\partial t} = & F [-k_1 n_{a0} n_{b0} (1 - s_1)] \\ & + \frac{\partial F}{\partial s_1} \{k_1 s_1 (1 - s_1) (n_{a0} + n_{b0} - 1) \\ & + k_2 [(s_1 - s_2) - n_{c0} (s_1 - s_2)]\} \end{aligned} \quad (30)$$

The boundary conditions are

$$F(0, 0; 0) = 1 \quad \text{and} \quad F(1, 1; t) = 1$$

The mean values of the numbers  $n_d$  and  $n_e$  of the elements  $D$  and  $E$  respectively are then as functions of time

$$m_{n_d}(t) = \left[ \frac{\partial}{\partial s_1} \log F(s_1, s_2; t) \right]_{s_1=s_2=1} \quad (31)$$

$$m_{n_e}(t) = \left[ \frac{\partial}{\partial s_2} \log F(s_1, s_2; t) \right]_{s_1=s_2=1} \quad (32)$$

The dispersions are

$$\sigma_{n_d}^2(t) = \left[ \frac{\partial^2}{\partial s_1^2} \log F(s_1, s_2; t) + \frac{\partial}{\partial s_1} \log F(s_1, s_2; t) \right]_{s_1=s_2=1} \quad (33)$$

$$\sigma_{n_e}^2(t) = \left[ \frac{\partial^2}{\partial s_2^2} \log F(s_1, s_2; t) + \frac{\partial}{\partial s_2} \log F(s_1, s_2; t) \right]_{s_1=s_2=1} \quad (34)$$

Thus the model of the complex may be represented as a black box with 2 inputs:  $k_1$  and  $k_2$ , with an initial state, characterized by the vector components  $n_{a0}, n_{b0}, n_{c0}, 0, 0$  at time  $t = 0$ , and with 2 outputs:  $m_{n_d}(t), m_{n_e}(t)$  or  $\sigma_{n_d}^2(t), \sigma_{n_e}^2(t)$ .

More generally, if  $k_1$  and  $k_2$  are not constants, set in advance, but change in time under the control action, the corresponding Markov process is non-homogeneous.

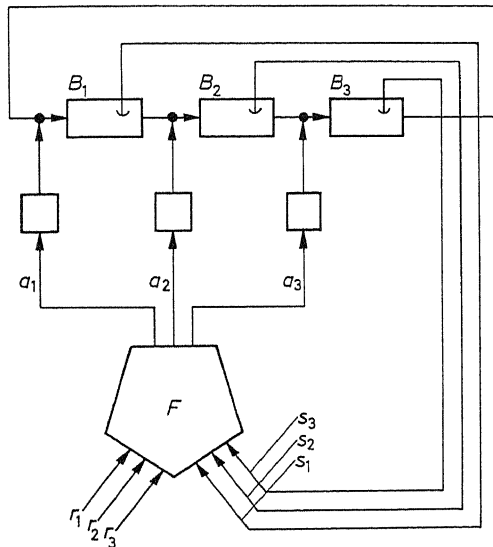


Figure 6

On Figure 6 there is a closed oriented chain of reservoirs  $B_1, B_2, B_3$ , containing many elements of the same type. This chain forms the complex. The acting variables of the formator  $a_1, a_2, a_3$  control the probability transducers represented by full points. The probabilities of the random transitions of elements from one reservoir to another are thus controlled. The aim of the function of the system is to reach a repartition of the elements over the reservoirs, prescribed by the command variables  $r_1, r_2, r_3$ . The total number of elements is  $N$ . The number of elements contained in the reservoir  $B_2$  at time  $t$  is  $x_2$ . The probability of the transition  $x_2 \rightarrow x_2 + 1$  in the time interval  $(t, t + \Delta t)$  is

$$\lambda_{x_2} \Delta t + o(\Delta t)$$

where, at first, let the  $\lambda$  in the relation

$$\lambda_{x_2} = \lambda \cdot x_2 \quad (35)$$

be constant in time.

If at time  $t$  the reservoir  $B_2$  contains  $x_2$  elements, the probability of the transition  $x_2 \rightarrow x_2 - 1$  in the time interval  $(t, t + \Delta t)$  is

$$\mu_{x_2} \Delta t + o(\Delta t)$$

where, at first, let the  $\mu$  in the relation

$$\mu_{x_2} = \mu \cdot x_2 \quad (36)$$

be constant in time.

The probability of the transition to a number of elements other than  $x_2 + 1$  or  $x_2 - 1$  is  $o(\Delta t)$ .

The probability of no change in the time interval  $(t, t + \Delta t)$  is

$$1 - (\lambda_{x_2} + \mu_{x_2}) \Delta t + o(\Delta t)$$

By making, in addition, similar assumptions for the reservoirs  $B_3$  and  $B_1$ , one gets the system of differential equations:

$$\frac{dP_{x_i}(t)}{dt} = \lambda_{x_i-1} P_{x_i-1}(t) - (\lambda_{x_i} + \mu_{x_i}) P_{x_i}(t) + \mu_{x_i+1} P_{x_i+1}(t),$$

for  $x_i = 1, 2, \dots$  and  $i = 1, 2, 3$  (37)

When the parameters  $\lambda$  and  $\mu$  in relations as (35) and (36) change in time, under the control action of the formator, one has to deal with Markov processes of the birth and death type non-homogeneous in time, describing the development of each of the reservoirs.

Another example of system may be suggested with complexes whose schematic representation is in the form of a two-dimensional array of zones, which may have, for example, a rectangular (Figure 7) or a triangular (Figure 8) form. The zones contain many elements. The transition of the elements from one zone to other zones is controlled by probabilistic transducers steered by the acting variables of the formator and represented as full dots. On Figure 7 the state of the selected zone  $R_{22}$  is a function of the states of the neighbouring zones. Considering at first a Markov process homogeneous in time as a model of the development of the zone  $R_{22}$ , which can be represented as a

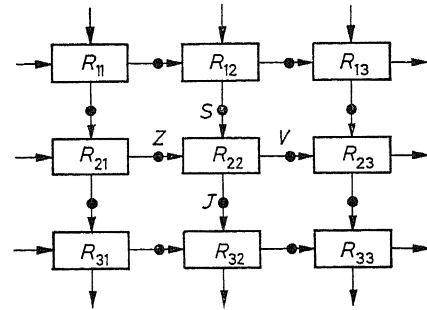


Figure 7

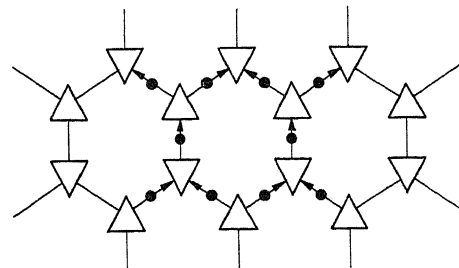


Figure 8

rectangle with two inputs and two outputs, one makes the following assumptions:

The number of elements in the zone  $R_{22}$  at time  $t$  is  $x_{22}$ .

The probability of the transition  $x_{22} \rightarrow x_{22} + 1$  in the time interval  $(t, t + \Delta t)$  is

$$\lambda_{x_{22}s} \Delta t + \lambda_{x_{22}z} \Delta t + o(\Delta t)$$

The probability of the transition  $x_{22} \rightarrow x_{22} - 1$  in the time interval  $(t, t + \Delta t)$ , if at time  $t$  the zone is in state  $x_{22}$  ( $x_{22} = 1, 2, \dots$ ), is

$$\mu_{x_{22}j} \Delta t + \mu_{x_{22}v} \Delta t + o(\Delta t)$$

The probability of the transition to a state other than  $x_{22} + 1$  or  $x_{22} - 1$  is  $o(\Delta t)$ .

The probability of no change of state is

$$1 - (\lambda_{x_{22}s} + \lambda_{x_{22}z} + \mu_{x_{22}j} + \mu_{x_{22}v}) \Delta t + o(\Delta t)$$

The corresponding Markov process pertaining to the zone  $R_{22}$  is of the birth and death type.

Because of the interrelation of the zones there is an interdependence between the parameters  $\lambda$  and  $\mu$  relative to neighbouring zones. It may be e.g.

$$\left. \begin{aligned} \lambda_{x_{22}s} &= \mu_{x_{12}j} \cdot x_{12}; & \mu_{x_{22}j} &= \lambda_{x_{32}s} \cdot x_{32} \\ \lambda_{x_{22}z} &= \mu_{x_{21}v} \cdot x_{21}; & \mu_{x_{22}v} &= \lambda_{x_{23}z} \cdot x_{23} \end{aligned} \right\} \quad (38)$$

The quantities  $\mu_{ikj}$  and  $\mu_{ikv}$ , where  $i, k = 1, 2, \dots, n$ , may be arranged into quadratic matrices. Owing to the action of the probabilistic transducers, the  $\mu_{ikj}$  and  $\mu_{ikv}$  change in time.

### Prospects of Development of the Theory and of its Applications

There is a large field for the development of the theory of systems with automatic control of configuration. The methods and results of the statistical mechanics form the basis for the dynamics of multivariable complexes. The representativeness of mathematical models is to be checked against physical measurements. The solution of problems, related to the Markov

processes of configurational development non-homogeneous in time, could be aided by modelling the process on simulators using random signal generators.

The field of application of the theory may be seen, for example, in these directions: the influencing of the formation of strips of molecules; the automatic control of the cultivation of microorganisms, such as algae; the formation of random nets.

*Fruitful suggestions from Professor Robert Fortet of Paris and from Professor Jaroslav Kožesník of Prague are gratefully acknowledged.*

### References

- MESAROVIĆ, M. D. *The Control of Multivariable Systems*. 1960. New York; The Technology Press of MIT and Wiley.
- HOWARD, R. *Dynamic Programming and Markov Processes*. 1960. New York; The Technology Press of MIT and Wiley.
- СОКОЛОВ, Н. Т. Пространственные матрицы и их приложения. Физматгиз, Москва, 1960
- SUGIMORI, MAKOTO. Binomial probabilistic sequential circuit. *Rev. Electr. Commun. Lab., Tokyo*, 9, Nos. 9-10 (1961) 627-654
- KOŽEŠNÍK, J. A simple stochastic model of continuous cultivation of microorganisms in several basins. *Second Internat. Symp. Continuous Cultivation of Microorganisms*, Prague, 18-23 (1962) 6
- JACOBSON, H. The informational content of mechanisms and circuits. *Inform. Control*, 2 (1959) 285-296
- BHARUCHA-REID, A. T. *Elements of the Theory of Markov Processes and their Applications*. 1960. New York; McGraw-Hill
- MØLLER, CHR. KN. Electrochemical investigation of the transition from tetragonal to cubic caesium plumb chloride. *Mat. Fys. Medd. Dan. Vid. Selsk., Copenhagen*, 32, No. 15 (1960)
- AUGENSTINE, L. G. Protein structure and information content, *Symposium on Information Theory in Biology*, pp. 102-123. 1958. London; Pergamon Press
- FOERSTER, H. v. On self-organizing systems and their environments. *Self-organizing Systems*. 1960. New York. Pergamon Press
- KOCHEN, M. Circle networks of probabilistic transducers. *Inform. Control*. 2 (1959) 168-182
- SINGER, K. Application of the theory of stochastic processes to the study of irreproducible chemical reactions and nucleation processes. *J. roy. Statist. Soc., ser. B*. 15 (1953) 92-106

### DISCUSSION

J. SKLANSKY, *RCA Laboratories, Princeton, New Jersey, U.S.A.*

The possibility of using Markov chain models to gain insight into configuration-controlled processes—or, equivalently, 'self-organizing' processes—is attractive, because the theory of Markov chains is so well developed. The examples in the paper demonstrate that these models offer a way of analysing processes involving large numbers of homogeneously distributed elements in one or a few enclosed volumes.

Unfortunately, these models do not seem so promising for random nets. In random nets we generally cannot find a useful division of the net into 'zones', as suggested in the paper, because in a random net the coupling among the zones usually will involve much more than simple neighbourhood coupling. This will be true, for example, in an 'elementary perceptron'<sup>1</sup>, in which each sensor unit ('S unit') is randomly connected to a subset of the set of associator units ('A units'), with no favouring of nearest neighbours.

Another obstacle to using Markov chain models for random nets is the high dimensionality of the resulting Markov chains. To demonstrate this, consider the very simple perceptron shown in Figure A. The reinforcement signal  $e$  adds an increment  $\eta$  to either or both of the gains  $k_1$  and  $k_2$ . If the stimulus is a sequence of statistically

independent vectors  $\{s_1(n), s_2(n)\}$ , and if we define the state of the system as the vector  $\{k_1, k_2\}$ , the Markov chain model of the system has the form of a two-dimensional random walk, as shown in Figure B.

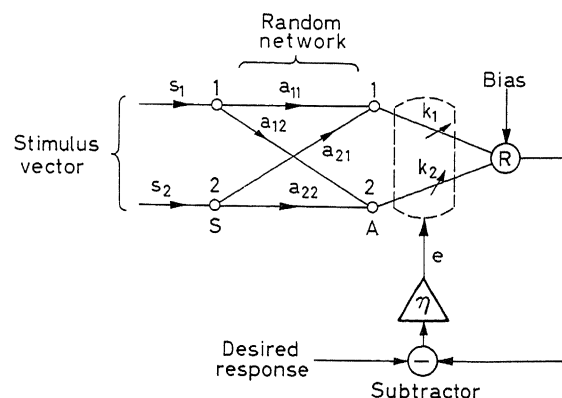


Figure A. A very simple perceptron with error-corrective reinforcement

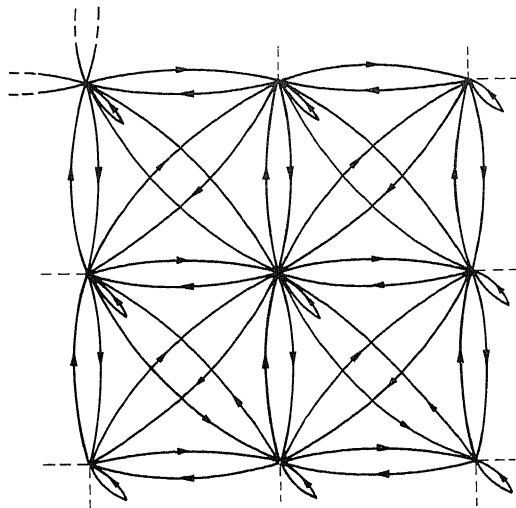


Figure B. A two-dimensional random walk

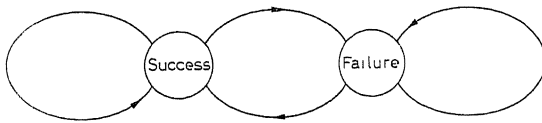


Figure C. A simple two-state model

If each  $k_i$  ranges over  $N$  values, the number of states in the Markov chain model is  $N^2$ . In more realistic perceptrons, however, the number of  $A$  units is large—100 or more. In general, the Markov chain model will be an  $\alpha$ -dimensional random walk with  $N^\alpha$  states, where  $\alpha$  is the number of  $A$  units. If  $N = 10$  and  $\alpha = 100$ —these are very conservative estimates—the stochastic matrix of the Markov chain obviously cannot be held in the memory of any real computer.

One approach which may overcome this difficulty is through a study of the eigenvalues of  $\alpha$ -dimensional random walks<sup>2</sup>. Another approach is to analyse only the convergence properties of the process as the time,  $n$ , approaches infinity<sup>3</sup>.

A third approach is to view the entire processes in the manner of mathematical psychologists toward human beings: i.e., partition of the externally observed behaviour patterns into a few states. A simple example of such a partitioning is shown in Figure C. Here the entire process is modelled by a two-state Markov chain, even though an internal examination of the system or 'organism' will most likely yield a many-state Markov chain. In the simple two-state model in the figure, one state corresponds to a 'successful' behaviour pattern; the other state to an 'unsuccessful' behaviour pattern.

We see, then, that a major problem in the study of self-organizing random nets is the modelling of certain classes of many-state Markov chains by few-state Markov chains. As yet, little attention has been given to this problem. But the interest in it will grow as the population of artificial self-organizing systems grows. With the advent of this new form of 'population explosion', engineers will take on more of the role of psychologists, and, we believe, will tend to describe the behaviour of self-organizing systems by few-state Markov chains.

#### References

- <sup>1</sup> ROSENBLATT, F. *Principles of Neurodynamics—Perceptrons and the Theory of Brain Mechanisms*. Rep. No. VG-1196-G-8, Office of Naval Research, Information Systems Branch, Washington 25, D.C., U.S.A. (March 1961)
- <sup>2</sup> MARCUS, M. Basic theorems in matrix theory. Superintendent of Documents, U.S. Government Printing Office, Washington 25, D.C., U.S.A. (January 22 1960)
- <sup>3</sup> KONHEIM, A. G. A geometric convergence theorem for the perceptron. *J. Soc. appl. Math.* Vol. 11, No. 1 (March 1963) 1–14

J. BENEŠ, *in reply*

I agree fully with Mr. Sklansky that the major problem in the study of self-organizing random nets is the modelling of certain classes of many-state Markov chains by few-state Markov chains. However, I would like to express my opinion on the direction of research, and to deal with both difficulties pointed out by Mr. Sklansky.

One way seems to be to consider the formator as a system which is in part organized in levels, these levels being hierarchically linked. In such a case it seems to me that the concept of the division of the complex into zones could be maintained and could be useful. Such a possibility can be seen, of course, if we consider random nets not specifically for perception but for the study of systems of organization, as in economical systems, distribution nets etc.

R. TARJAN, *1, Ponty-utca 2, Budapest, Hungary*

The problem of configuration-controlled or self-organized systems is indeed a very exciting one, and rightly deserves great interest. The introduction of probabilistic transducers as a method of forming zones in random nets seems also a necessary step in attaining a model of a living organism such as the central nervous system. However, looking at the configuration controlled systems as a possible mathematical model of the central nervous system, it seems to be necessary, that besides the five basic operations introduced by the author, a sixth operation be introduced, namely that of the separation. This would be clearly an analogon of the biological inhibition which, as is well known, plays a dominant role in the proper functioning of the brain, e.g. in the selection of memory contents.

One further remark concerns the measure of order given by the author in (2). This measure takes into account only the numbers of the effective respective possible connections. Now a complex such as the central nervous system is not only connected, but also has a very definite organization necessary for survival. Intuitively it is clear that the organization is a selected set of connections, the selection being made so as to suit a given task. It is also intuitively clear that, given the task, there are many possible organizations, one of them being more complicated than the others. So, in dealing with models of such complexity, the problem of having rigorous definitions of the notations of organization arises, such as the complexity of organization with respect to a given task. The situation is analogous to an optimizing system which can be optimized only with respect to a well-defined criterion. I would be obliged if Professor Beneš will kindly comment on this problem.

J. BENEŠ, *in reply*

Dr. Tarjan's first question concerns the selection of zones. I think this is a question of applying some mathematical statistical methods to select the number and the place of the zones which will be used for measuring the state of the complex as a whole, to have zones selected which would be really representative.

The first method is mentioned in the paper. A second method which now seems very promising is the use of the well-known technique of fibre-optics, and using this technique it would perhaps be possible to solve this problem of selection of zones.

The second question is whether or not to use redundancy in the measuring of organization. I would say that the concept of configurational redundancy seems to be important from the point of view of theory. In order to be able to use certain concepts of information theory in my paper I have already pointed out that there are very simple and, for technical use, important criteria of ordering and measuring; for instance, the number of the occupied sites in a lattice or the number of elements of a certain kind and type which had been formed by reaction.

# Approximation of Industrial Control Systems for Optimized Non-linear Control with Restricted Rate of Correction Using Conventional Controllers

E. PAVLIK

## Summary

The positioning speed of servo-motors for industrial process control systems is always limited. With control loops of low inertia this speed limitation has a high influence on the control behaviour and causes a considerable non-linearity of the control loop. According to a theory<sup>1-3</sup> for systems with limited parameters it has been proved that the optimum transient response for a discrete disturbance and the optimum transient responses for different disturbances are predetermined. Due to the complexity and sensitivity to disturbances of the required control equipments, however, the experiences gained from the optimization theory have been of little consequence for practical control problems. Thus no rules for design, adjustment and quality of conventional continuous controllers of the *PI* or *PID* type, providing a good approximation to the above transient response characteristics, have been formulated. The same applies to the three-position controller (pulse controller) commonly used in practice. For this reason, attempts are being made to develop methods giving suitable approximations to the optimum characteristics. For simple systems, the adjustment parameters of the above-mentioned industrial controllers can be calculated by an approximation method.

## Sommaire

La limitation des vitesses de positionnement des servomoteurs industriels, dans des boucles de commande à faible inertie, influence grandement le comportement de la commande et, souvent, produit une non-linéarité considérable. Dans certaines études<sup>1-3</sup>, on a prédéterminé la réponse transitoire optimale par rapport à une perturbation discontinue et à un ensemble de perturbations diverses. Toutefois, l'équipement nécessaire est à la fois trop complexe et trop sensible aux perturbations. De ce fait, théorie d'optimisation n'a que peu d'effets dans la pratique où l'on utilise des régulateurs classiques du type *PI*, *PID*, à trois positions, etc. Dans ce rapport, on présente une méthode de calcul approximative, permettant de définir les valeurs optimales des paramètres réglables des régulateurs mentionnés ci-dessus.

## Zusammenfassung

Die Stellgeschwindigkeit des Stellantriebs technischer Regelanordnungen ist stets begrenzt. Besonders bei schnellen Regelkreisen ist diese Begrenzung der Stellgeschwindigkeit von großem Einfluß auf das Regelverhalten. Die Theorie<sup>1-3</sup> für Systeme mit begrenzten Parametern liefert den Beweis, daß für eine bestimmte Geschwindigkeitsbegrenzung sowohl der optimale Übergangsprozeß bei einer diskreten Störung als auch die Gesamtheit der optimalen Übergangsprozesse bei verschiedenen Störungen bereits festliegen. Wegen der Kompliziertheit und Störanfälligkeit der dazu erforderlichen Regelorgane haben sich die von der Optimaltheorie vermittelten Erkenntnisse bis jetzt wenig befruchtend auf die regelungstechnische Praxis ausgewirkt. So existieren z. B. keine Vorschriften für den Entwurf, die Einstellung und die Beurteilung von konventionellen kontinuierlichen Reglern des *PI*- oder *PID*-Typs, welche eine möglichst gute Annäherung an den Optimalvorgang bzw. die Gesamtheit der Optimalvorgänge gewährleisten. Das gleiche gilt für die in der regelungstechnischen Praxis weitverbreiteten

quasistetigen 3-Punkt-Regler (Impulsregler) mit nachgeschaltetem Stellmotor. Deshalb wird versucht, Methoden zu entwickeln, die brauchbare Näherungen an die Optimalvorgänge ermöglichen sollen. Für einfache Systeme lassen sich die Einstellparameter der obengenannten technischen Regler zur Realisierung optimaler Vorgänge mit Hilfe einer Näherungstheorie berechnen.

## Theory of Optimized Systems

The theory of optimum response of a system with one or several limited parameters to varying external influences has been fully established in the previous studies<sup>1-3</sup>. It is not within the scope of this contribution to repeat this theory in all details. However, the knowledge of the theory regarding speed restricted systems is indispensable for the understanding of the following statement. For this reason, a simplified summary of the essentials is given here:

Let us assume a coordinate  $x_i$  of a transfer system  $F$  (e.g. the differential quotient in time of the correcting action) is restricted to a given interval.

$$x_{i1} \leq x_i \leq x_{i2} \quad (1)$$

In order to obtain the fastest possible change of the output  $x$  of the system, the coordinate  $x_i$  has one of the two limit values which will control the system in the desired sense. The value of the output  $x$  and of its derivatives is determined by the initial state of the system  $x_A, \dot{x}_A, \dots$  the coordinate  $x_{i1}$  or  $x_{i2}$  respectively and the time  $t$  elapsed from the beginning of the process up to the particular instant  $t$ .

$$x = f_0(x_A, \dot{x}_A, \ddot{x}_A, \dots, x_{i1} \text{ or } x_{i2} \text{ respectively}; t)$$

$$\dot{x} = f_1(x_A, \dot{x}_A, \ddot{x}_A, \dots, x_{i1} \text{ or } x_{i2} \text{ respectively}; t) \quad (2)$$

$$\ddot{x} = f_2(x_A, \dot{x}_A, \ddot{x}_A, \dots, x_{i1} \text{ or } x_{i2} \text{ respectively}; t)$$

$$\vdots$$

Since the limited coordinate  $x_i$  is constant at the instant when  $x$  reaches a given value (e.g. the desired value), the derivatives of  $x$  will generally differ from the derivatives of the given value, so that no steady state of the system is obtained. In order to achieve a steady state at the end of the transient response, the value of the restricted coordinate must vary in a definite manner. The system responds quickest if the steady state is produced by keeping  $x_i$  in alternating steps at one of its limit values, provided that, besides the restriction of coordinate  $x_i$ ,



the transfer function  $F$  is linear and has real non-positive poles<sup>1-3</sup>.

To this end, the process is split up into a number of intervals  $t_1, t_2, t_3, \dots$ . During each interval the restricted coordinate  $x_i$  is maintained constant at one of the two limit values. When changing over to the next interval,  $x_i$  is given the opposite limit value, so that it alternates from one limit value to the other.

The order of the (linear) transfer differential equation, which relates the output  $x$  and its derivatives with the restricted coordinate, is indicated by  $n$ . The initial condition and the desired condition are determined by the output  $x$  and its  $n-1$ th derivative.

If  $F$  is, for instance, a guided system and the governing variable is  $Z_1$ , eqn (2) becomes:

$$\left. \begin{aligned} x &= Z_1 = f_0(x_A, \dot{x}_A, \dots, t_1, t_2, \dots) \\ \dot{x} &= \dot{Z}_1 = f_1(x_A, \dot{x}_A, \dots, t_1, t_2, \dots) \\ &\vdots \\ &\vdots \\ &\vdots \\ x &= Z_1 = f_{n-1}(x_A, \dot{x}_A, \dots, t_1, t_2, \dots) \end{aligned} \right\} \quad (3)$$

(3) is a family of  $n$  equations and, in order to satisfy the  $n$  conditions, at least  $n$  independent parameters must be available, e.g. the durations of the intervals  $t_1, t_2, \dots$ . In this case,  $n$  is the minimum number of intervals required to bring the system from the initial condition to the desired condition. The family of equations has, therefore,  $n$  unknowns,  $t_1, t_2, \dots, t_n$ , the  $n$  intervals, which can be determined as the roots of the family of eqns (3).

Equation (3) can also be regarded geometrically. Considering  $Z_1$  as time function  $Z_1(t)$ , eqn (3) represents the necessary and sufficient condition to accomplish that, after the last interval  $t_n$  has elapsed, the output  $x(t)$  is identical with the input  $Z_1(t)$ , because  $x(t)$  and  $Z_1(t)$  are coincidental functions up to the  $n$ th order, provided the condition of reproducibility [eqn (4)] is fulfilled.

This is the case if

$$[a_n Z_1^{(n)} + a_{n-1} Z_1^{(n-1)} + \dots + a_1 \dot{Z}_1 + a_0 Z_1] \leq K \quad (4)$$

where  $K$  is the maximum value of the restricted coordinate. Furthermore, the differential equation of the transfer system  $F$  must have the parameters  $a_n, a_{n-1}, \dots, a_0$ .

If, for example,  $Z_1$  is a step function of the amplitude  $Z_{11}$ , the  $n$  optimum switching instants  $t_i$  ( $i = 1 \dots n$ ), corresponding to the  $n$ th order of the system, can be calculated from eqn (3). The switching instants  $t_i$  are functions of the system parameters and of the step magnitude  $Z_{11}$ . Generally, for an input  $Z_1$ , the switching instants are functions of the system parameters and of the parameters of  $Z_1$ . If within the intervals found in this way the restricted coordinate is alternately kept at one of its two limit values, the variable  $x$  will reach the desired condition aperiodically and in the shortest possible time. Optimum switching instants are frequently obtained as a result of automatic control actions by feeding back the output  $x$  to the input of the restricted coordinate.

At this juncture it is of particular importance to introduce the term 'completeness of the optimum processes', formulated by Solodovnikov<sup>3</sup>.

If a system is said to qualify for 'completeness of optimum processes', it means that optimum switching instants and thus

optimum responses are automatically achieved not only for one discrete external influence, but optimum transfer conditions exist in such a system for a whole family of inputs, e.g. step disturbances of any amplitude or ramp disturbances of any slope, or both kinds of disturbances at the same time. This is much the same as with linear systems where, once the correct adjustment for a certain kind of input has been found, this will also hold good for any amplitude. This amazing fact can be explained as follows:

Let us assume that for an arbitrary external influence  $Z$  (variation of desired value or load change) the optimum switching instants have been calculated from eqn (3). As pointed out above, these switching instants are functions of  $Z$ :

$$\left. \begin{aligned} t_1 &= \varphi_1(Z) \\ t_2 &= \varphi_2(Z) \\ &\vdots \\ t_n &= \varphi_n(Z) \end{aligned} \right\} \quad (5)$$

The transient responses of the output  $x$  and its derivatives are functions of these intervals:

$$\left. \begin{aligned} x(t) &= f(t_1, t_2, \dots, t_n) \\ \dot{x}(t) &= \dot{f}(t_1, t_2, \dots, t_n) \\ &\vdots \\ &\vdots \\ x &= f(t_1, t_2, \dots, t_n) \end{aligned} \right\} \quad (6)$$

Elimination of all interval times from the equation system (6), would enable relationships to be established between  $x$  and its derivatives, independent of the interval times. If it is, for example, assumed that it is possible to eliminate all interval times from the equations for  $x(t)$  and  $\dot{x}(t)$ , as can indeed be done in the case of some simple systems with practical applications, the equation takes the form

$$x = \varphi(\dot{x}) \quad (7)$$

and is, therefore, not dependent on the interval times. Equation (7) can be re-arranged to

$$x - \varphi(\dot{x}) = 0 \quad (8)$$

The system yields the output  $x$ , which is transformed to  $\varphi(\dot{x})$ . The latter is subtracted from  $x$  in accordance with eqn (8) and the difference is fed back, as a compensating signal, to the input, namely the summing point of the external influence or the restricted coordinate respectively. Obviously, this will produce an effect only if the process is not an optimum. Expressed in terms of control technique, the purpose of the feedback is to bring the actual process nearer to optimum conditions, since for the optimum process the relation (8) is under all circumstances obligatory.

Frequently it is not possible to eliminate completely the interval times from (6) without involving higher derivatives. In this case, the application of the function transfer  $\varphi(\dot{x}, \ddot{x}, \dots)$ , found from eqn (8) becomes impracticable, owing to the incompatibility of higher derivatives with the noise level. Furthermore, higher derivatives are mostly variables of transcendental functions, which cannot be made use of in practice at a reasonable expense.

## Theory of Quasi-Optimum Systems

### Approximation via the Optimum Switching Instants

If the difference between the desired value and the actual value is  $x_w$ , the  $n$  optimum switching instants  $t_1, t_2, \dots, t_n$  can be determined by the following rearrangement of the family of eqn (3):

$$\begin{aligned} 0 &= x_w(t_n) = f_0(x_A, \dot{x}_A, \dots, x_A^{(n-1)}, t_1, t_2, \dots, t_n) \\ 0 &= \dot{x}_w(t_n) = f_1(x_A, \dot{x}_A, \dots, x_A^{(n-1)}, t_1, t_2, \dots, t_n) \\ &\vdots \\ 0 &= x_w^{(n-1)}(t_n) = f_{n-1}(x_A, \dot{x}_A, \dots, x_A^{(n-1)}, t_1, t_2, \dots, t_n) \end{aligned} \quad (9)$$

Assume that the  $t_1 \dots t_n$  have been determined for one particular external influence  $Z$  (variation of the desired value or load change). For a system of the  $n$ th order the  $t_i$  ( $i = 1 \dots n$ ) are functions of the external influence and, furthermore, according to eqn (9), of the initial values  $x_A^{(i)}$  [ $i = 0 \dots (n-1)$ ] and of the balance values  $x_w^{(i)}(t_n)$  [ $i = 0 \dots (n-1)$ ].

Now assume that the correcting servo-motor together with the controlled plant form the system of the  $n$ th order and that the correction speed of the servo-motor is limited. This speed-limited servo-motor element with the maximum speeds of operation  $+K_0$  (forward) and  $-K_0$  (reverse) can be considered as an integrator with integration time constant 1, preceded by a restricting element, having the saturation values  $+K_0$  and  $-K_0$  (Figure 1).

It follows from eqn (9) that the restricting or relay element, preceding the motor, must change  $(n-1)$  times in the intervals  $t_1 \dots t_n$  from the one extreme position to the other, in order to correct the deviation  $x_w$  and all its  $(n-1)$  derivatives within the shortest possible time.

Figure 2 shows the block diagram of a system incorporating a controller  $R$  of a certain type [e.g., a conventional continuous proportional-plus-integral (PI) or proportional-plus-integral-plus-derivative (PID) controller] which may have  $K$  free setting parameters.

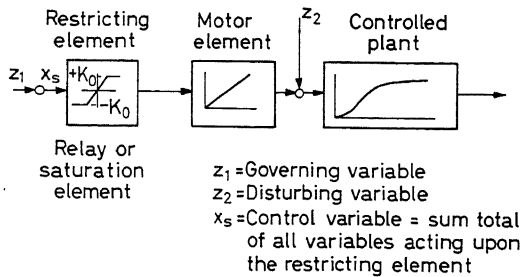


Figure 1

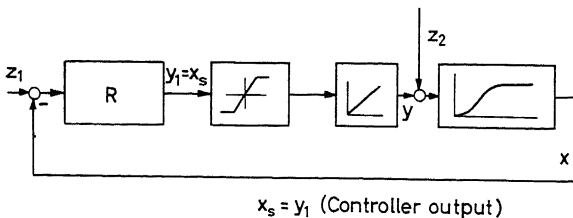


Figure 2

If the correcting servo-motor has a static proportionality effect (steady-state offset caused by rigid feedback), the block diagram of the system will slightly change as indicated in Figure 3. The controller may again have  $K$  free setting parameters.

The block diagram, Figure 4, represents an impulse-type system with a 2- or 3-position relay as control amplifier.  $Ru$  is the feedback, having again  $K$  free setting parameters.

Figures 3 and 4 are typical block diagrams for a large number of (continuous or quasi-continuous) control systems met in practice.

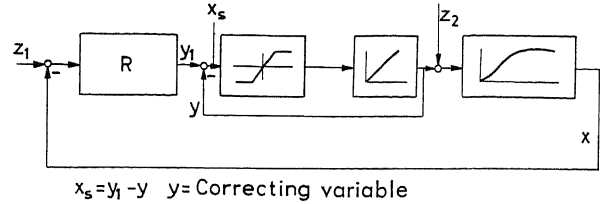


Figure 3

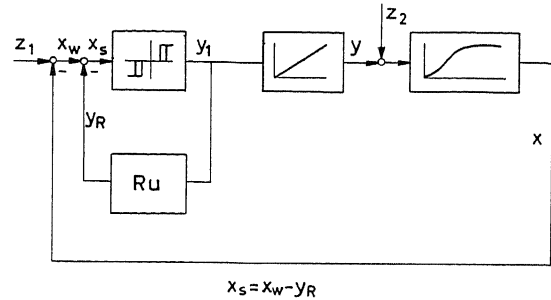


Figure 4

The input amplitudes required for moving the restricting elements to the end positions are insignificant and can be neglected. The condition for obtaining optimum response can be written as

$$x_s(t_i) = 0 \quad \text{for } i = 1, \dots, n \quad (10)$$

because the servo-motor has to be changed over at each instant  $t_1 \dots t_n$ . However,  $x_s$  must not disappear between switching instants  $t_i$ , so that the restricting element can definitely reach the extreme positions.  $x_w$  can be calculated for any instant  $t_1 \dots t_n$ .

Since it has been assumed that the controller in Figures 2 and 3 has  $K$  free setting parameters, the output  $y_1$  of the controller (transfer function) will be:

$$y_1 = f(K_1, K_2, \dots, K_K, x_w, t) \quad (11)$$

For conventional continuous controllers, for instance:

P controller:  $y_1 = K_1 x_w$

PI controller:  $y_1 = K_1 x_w + K_2 \cdot \int_0^t x_w dt$

PID controller:  $y_1 = K_1 x_w + K_2 \cdot \int_0^t x_w dt + K_3 \dot{x}_w$

For the controller in *Figures 2* and *3* the number of free setting parameters  $K$  must equal  $n$  in order to satisfy eqn (10). One can, therefore, write down, e.g., for the system in *Figure 2*:

$$\begin{aligned} x_s(t_1) &= y_1(t_1) = 0 = f(K_1, \dots, K_n, t_1) \\ x_s(t_2) &= y_1(t_2) = 0 = f(K_1, \dots, K_n, t_1, t_2) \\ &\vdots \\ x_s(t_n) &= y_1(t_n) = 0 = f(K_1, \dots, K_n, t_1, t_2, \dots, t_n) \end{aligned} \quad (12)$$

A similar equation system can be developed for that in *Figure 3*:

$$x_s(t_i) = (y_1 - y)_i = 0 \quad \text{for } i = 1, \dots, n \quad (13)$$

in which way  $n$  equations are again obtained for the determination of the  $n$  required parameters  $K$ .

It has been assumed that the feedback of the system shown in *Figure 4* has again  $K$  free setting parameters. In this case, also,  $K$  must equal  $n$  in order to satisfy eqn (10) and to obtain the optimum response for one disturbance  $Z_1$  or  $Z_2$ .

Summing up, it can be stated that, in order to obtain the optimum response to a particular disturbance  $Z$ , the arrangement of a control system of the  $n$ th order (including the correcting servo-motor), must be such that at least  $K = n$  free setting parameters are available (with the only exception of a controlled plant of order 0). The response to any other disturbance will not be the optimum. In the case of conventional controllers without non-linear computer elements, an approximation to the 'completeness of the optimum processes' (which can be accomplished only by non-linear elements) can be obtained by determining the optimum parameters for two or several disturbances and by interpolating between them.

Attention must be drawn to one important fact emerging from the equation systems for the determination of the  $K$  parameters: since a real controller has hardly more than three free setting parameters, it is not possible to optimize completely systems of more than the third order. However, the present contribution deals particularly with fast control systems, which generally belong to the first and second order, so that one need not worry about the above-mentioned limitation.

#### Approximation via the Optimum Control Function

The following relationship between the output  $x$  and its derivatives exists in an assumed optimized system which satisfies eqn (9).

$$x = \varphi(x, \dot{x}, \ddot{x}, \dots) \quad (14)$$

The optimum control function  $\varphi$  is normally non-linear. As mentioned above, in order to obtain 'completeness of the optimum processes', the control loop must comprise a function couverter, which is responsible for satisfying the relationship

$$x - \varphi(x, \dot{x}, \ddot{x}, \dots) = 0$$

at all switching instants  $t_i$ .

If this function couverter is too costly, or, in case it is necessary, with a view to suppressing small disturbance oscillations, to neglect the higher derivatives or powers of derivatives, an approximation to the relationship (14) must be attempted with the means at our disposal. The same applies to conventional  $P$ ,  $PI$  and  $PID$  controllers, which are themselves primitive computers. A generally applicable approximation to the optimum control function cannot be given. An approxima-

tion cannot be successful, unless the influence of important variables is adequately taken into consideration. For instance, with simple systems, a close approximation can be achieved by substituting a linearized element of the form  $K_2 \dot{x}$  for frequently appearing elements of the optimum control function in the shape  $|\dot{x}| \dot{x}$ . It is frequently possible to approximate important control variables, particularly the higher derivatives, by the use of relatively simple means.

#### Conditions Under Which Quasi-optimized Systems Can Find Practical Application

If it is intended to approximate by determining the  $K$  free setting parameters of the system through the knowledge of the optimum switching instants, it must be borne in mind that the resulting values for the parameters can be practically used only if they are physically achievable and if their use would not impair the stability of the system. For instance, if eqn (10) yields a negative time constant for a feedback, this is obviously useless for practical purposes. It is, furthermore, obvious that if a calculation of this kind results in a negative value for the total gain of a feedback network, this makes no sense at all, since the transmittance of a continually negative feedback signal to the system input would lead to monotonous instability (unbalance of the control amplifier). The same applies, of course, to negative proportional bands, etc. Furthermore, the control functions of the approximation system, resulting from the determination of the  $K$  free setting parameters, must be such that the restricting element is kept to its extreme positions during the switching intervals; in other words, the control function  $x_s(t)$  of the approximation system must not disappear between the switching instants. This condition can be formulated by:

$$\begin{aligned} x_s(t) &\neq 0 & \text{for } t < t_n \\ &\text{and } t &\neq t_1, t_2, \dots, t_n \end{aligned} \quad (15)$$

$t_n$  = instant of the last switching operation because according to eqn (10):

$$x_s(t_i) = 0 \quad i = 1, \dots, n$$

Moreover, the deviation from 0 must be sufficient to secure that the limit values are definitely reached, which can be expressed by:

$$\frac{d[x_s(t)]}{dt} \neq 0 \quad \text{for } t = t_1, t_2, \dots, t_{n-1} \quad (16)$$

For the stability of the nil positions of an approximation system consisting of linear transfer elements except the restricting elements, the conditions previously outlined<sup>1</sup> are generally applicable. To comply with these conditions, the limit system, that is the system with an infinitely steep, unrestricted 'limiting element', must be stable and the transfer function  $U(t)$  of the linear part must satisfy the following requirements:

$$\begin{aligned} &\text{for } u(t=0) = 0 \\ &\dot{u}(t=0) > 0 \\ &\text{and for } u(t=0) = \dot{u}(t=0) = 0 \\ &\ddot{u}(t=0) > 0 \text{ and } \ddot{\ddot{u}}(t=0) < 0 \end{aligned} \quad (17)$$

This also means that the difference between the degrees of the denominator and of the numerator belonging to the transfer

function of the linear system part must be  $\leq 2$ . The following discussions which supplement the full treatment given to the subject<sup>1, 3</sup> will be of interest.

Investigations with an analogue computer have proved that with a difference between the degrees of denominator and numerator equal to 2 and compliance with the other conditions, the stability is secured in principle, however, that the system becomes considerably underdamped when the correcting time of the motor is of about the same order of magnitude as the total time lag of the controlled plant, which is just the range of interest for this contribution. The performance improves considerably if it is possible to reduce the difference between denominator and numerator to 1. The system then obtains the capacity of assuming sliding states (explained in reference 1) which is particularly favourable for the performance of systems governed by a variable.

A correcting servo-motor has a finite correcting speed and also a limited correcting action. This constitutes a further considerable non-linearity and a serious drawback for the quasi-optimization of systems, whether the approximation is attempted *via* the optimum switching instants or the optimum control function. Equation (3) allows a decision to be made immediately after calculation of the optimum switching instants, whether the limitation of the correcting action will be unfavourable to the chances of obtaining quasi-optimization. If the system is in equilibrium when the disturbance occurs, the correcting element assumes a certain position within its travel range. From this position of equilibrium, it will move towards the top or the bottom position, in accordance with the sign of the disturbance. Obviously, the first switching interval  $t_1$  must not produce a movement greater than the available remainder of the travel range, which is called  $h$ , otherwise the travel would be forcibly and prematurely terminated by the extreme position stop. If  $T_M$  is the running time of the motor over its whole travel range, the condition just outlined, for unhindered optimum response, can be written

$$t_1 \leq h \cdot T_M \quad (18)$$

Since  $t_1$  is always a function of the system parameters and disturbance amplitudes, the disturbance amplitudes to be expected have to be taken into account, which most probably involves a statistical problem. Studies carried out by means of an analogue computer have proved that the effect of the travel limitation mentioned above, dependent on the choice of the disturbance parameters, can considerably deteriorate the system response, however, without jeopardizing the stability.

Naturally, the theoretically developed quasi-optimized system must be a practical proposition with regard to instrumentation and must also conform with established control engineering requirements. Complex non-linear functions must be excluded immediately; this applies also to demands for accurate higher derivatives of the controlled condition, demands which, as experience has taught, are incompatible with the noise level of the system. This obstacle, however, can be overcome by the ingenious methods of originating feedback signals respectively prior to the controlled plant or the servo-motor.

## Conclusions

From the theoretical discussions and the investigations with an analogue computer, directions can be deduced for the further development of systems and instruments suitable for the

design of fast-acting automatic controls. These directions call for conceptions, which differ from those accepted up to now, or, at least, for modifications of such conceptions, particularly with reference to fast-acting automatic process control devices.

At present, almost without exception, *PI* controllers (mostly with fixed proportional band) are applied for fast control, particularly flow control, in conjunction with a servo-motor having restricted correcting speed. It has become apparent from the deliberations outlined above that this cannot be the correct solution. If, however, this arrangement has been chosen the controller should be given a derivative action, thus considerably narrowing the proportional band, particularly with a controlled plant of the first order. The closer approximation to the optimum control function thus achieved will result in a greatly improved quality of control.

It has been realized that it is advisable, for fast control systems, to abandon the combination of a linear *PI* or *PID* controller with a speed-restricted servo-motor element altogether and to pay more attention to relay systems. It has, furthermore, been shown that the use of a three-position relay instead of a continuous controller, the saturation values of which are equal to the response values of the relay, has little influence on the transfer behaviour of the system. It follows that a pneumatic or hydraulic system conforming to block diagram, *Figure 4*, will be simpler in instrumentation and better in control performance than a system such as that indicated in *Figure 3*.

It has, furthermore, been proved that a forced linearization of the correction transfer function of speed restricted servo-motors is inadvisable. In practice, linearization is achieved by using a positioner of the *PI* type for the regulation of the movement of the valve stem thus obtaining correction transfer functions having approximately equal time constants. If, after all, a positioner is employed, the restricted co-ordinate, that is the maximum available rate of correction, should be utilized to its full extent. The ensuing considerable non-linearity of the correcting action should not be corrected by additional devices fitted to the servo-motor but by adjustment of the controller to the particular non-linearity.

With reference to the conception of the 3-position controller the following can be said: up to now, the delaying feedback, typical for such systems, is normally designed in such a way that, with the feedback amplification remaining constant, adjustable time constants are available for the charging and discharging of the feedback capacitors. This produces an additional adjustability of the so-called integral (*I*) part of the controller. Parameter calculations and analogue studies have proved beyond doubt that this conception meets the requirements of a system with restricted rate of correction only to a small extent. For a good approximation to the 'completeness of the optimum processes' it is not necessary to have unequal charging and discharging time constants, but an adjustable feedback amplification is of paramount importance. In order to clarify this question with absolute certainty it was decided to carry out the analogue studies with a linear *RC* feedback and also with a non-linear feedback having unequal charging and discharging time constants, the feedback amplification remaining constant (as is the practice of many manufacturers). In all cases it was confirmed that it was not possible by this method to eliminate, even approximately, the unfavourable effect caused by the constant feedback amplification.

Finally, it should be stated that this paper has not, to any degree, exhaustively dealt with all the problems connected with the logical and reasonable design of quasi-optimized systems or for systems with restricted parameters. The practically always existent non-linearity of the controlled plant had, for instance, to be neglected. The most important of these non-linearities is represented by the non-constant static transfer behaviour (gain of the controlled plant). In order that this property may not pass entirely unnoticed, calculations have been carried out for various plant gains, with the result that

changes of about a factor of 2 do not have much effect on quasi-optimized systems or on their adjustable parameters.

## References

- 1 TSYPKIN, YA. S. *Theorie der Relaisysteme der automatischen Regelung*. München; R. Oldenbourg, 1958
- 2 FELDBAUM, A. A. *Optimale Prozesse in Regelsystemen*. München; R. Oldenbourg, 1962
- 3 SOLODOVNIKOV, W. W. *Grundlagen der selbsttätigen Regelung*. Vol. 2. München; R. Oldenbourg, 1959

## DISCUSSION

YA. Z. TSYPKIN, *Institute of Automation and Telemechanics, Kalanchevskaya 15-A, Moscow I-53, U.S.S.R.*

The problem of approximate realization of optimal (with respect to speed) processes by standard controllers is very important from the practical point of view, and in this field the author has obtained very interesting results.

As the optimal system operates in the large deviations mode, the condition of steady-state stability ('in the small') would not be enough in all cases. The steady state of a relay system should be stable 'in the large'. It is easy to write down the stability condition 'in the large' in the form:

$$\operatorname{Im} W^*(j\omega) < 0, \quad 0 \leq \omega < \infty \quad (1)$$

where  $W^*(j\omega)$  is the frequency response of the linear part of the relay system. This condition includes the conditions (17). Perhaps, with the exact realization of optimal processes, the stability conditions 'in the large' will be satisfied. In the approximate realization, instead of conditions (17), it is necessary to use condition (1).

These remarks are additional to the very rational and efficient way chosen by the author for the approximate realization of optimal systems by comparatively simple technical means.

E. PAVLIK, *in reply*

Many thanks to Professor Tsypkin for his very interesting contribution. It is true that condition (1) of his discussion remarks contains condition (17). As to exact optimum processes, eqn (17) is, in my opinion, sufficient, since the system at the last switching moment  $t_u$  comes automatically in the state of stability of small deviations, the controlled condition  $x_w$  and all its derivations being eliminated as far as the  $(n-1)$ th order. I agree with the Professor that condition (1) must be fundamentally complied with for the approximation system. For the studies carried out in analogue computation it is true that eqn (17) has proved sufficient. Professor Tsypkin's contribution is a valuable completion of my work.

F. SCHREINER, *Fuchshohl 18-20, Frankfurt/Main, Germany*

In his conclusions Dr. Pavlik proposes to abandon the combination of linear *PI* or *PID* controllers with a speed-restricted servo-motor for fast control systems, particularly flow control systems. He advises that more attention be paid to relay systems.

(1) What is the definition of a 'fast' control system, where a controller, as indicated in Figure 4 of the paper, gives better control results than a linear controller. In my opinion the quality of control depends on the ratio  $\frac{T_m \cdot y_z}{T_s}$ , where  $T_m$  is the running time of the motor for full range,  $T_s$  the time constant of the plant, and  $y_z$  the movement of the motor which is required to compensate for a disturbance of amplitude  $z$ .

(2) The positioning speed for an error amplitude  $x_w$  is proportional to  $\frac{1}{x_p \cdot T_n}$ . With higher order plants, the influence of the restricted positioning speed decreases, as  $\frac{1}{x_p \cdot T_n}$  must decrease.

(3) The author proposes to optimize for step disturbances. The results of his deliberations are not applicable to sinusoidal disturbances. Tests on the analogue computer have shown that the frequency range, in which the disturbance amplitudes are decreased by control, is narrowed with increasing running time. As Westcott pointed out, for linear control systems, a better control quality in one frequency range is obtained only at a cost of less quality in another. By a restriction in the closed loop the control quality in a certain frequency range will be adversely influenced. There will be no improvement in any other frequency range.

(4) Changes of the plant gain by the factor 2 will have no effect on quasi-optimized systems and their adjustable parameters only if the running time of the motor is longer than the time constant of the plant. To get better control results, the running time must be shorter. In this case the variation in plant gain by the factor 2 will eventually cause instability, as is known from the setting rules by Ziegler and Nichols.

E. PAVLIK, *in reply*

(1) Quick control systems are understood to be systems with the total delay time of the controlled plant being of the order of the time necessary for the servomotor to run through the total servo-system with a top floating speed. In the studies of analogue computation carried out, the relation  $T_m/T_s$  is about 1/3 to 3.

(2) I agree with this contribution but not with the linear interpretation in connection with the equation  $\frac{1}{x_p \cdot T_n}$ . The influence of the limited floating speed decreases with the increasing order of the controlled plant, the total delay time of the controlled plant logically increasing with the same.

(3) This point has arisen because of a translation error in the preprint, which has now been corrected. I agree, of course, with Mr. Schreiner in that the time lag of a servomechanism is not sufficient to correct the total control behaviour of the systems observed.

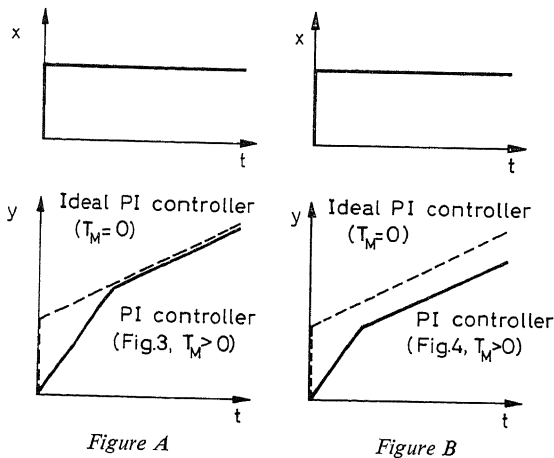
(4) The adjustment rules of Ziegler and Nichols refer to the linear systems which do not exist here. We are just starting from the fact that the floating speed reaches the limiting values.

W. BÖTTCHER, *Wolfgangstr. 56, Frankfurt/Main, Germany*

In his paper Dr. Pavlik has given interesting proposals for approximation methods for optimal characteristics of industrial controllers with speed-limited servomotors. To this the following remarks are added:

(1) In certain controller arrangements with speed-limited actuators the quality of control decreases compared with a linear *PID* controller, if the positioning speed required is faster than the maximum speed. Due to restricted positioning speed the control system (Figure 4) is damped with increasing disturbance amplitudes  $z$ , whereas the damping ratio decreases if a controller with a positioning servomechanism

with restricted speed (Figure 3) is used. Therefore, one has to optimize the system of Figure 3 for small disturbances and the system of Figure 4 for large disturbances. If this is not done the transient responses of Figures A and B are the result. The setting of the controllers and the speed of the motor are the same for the controller arrangements of Figures 3 and 4. The result of restricted speed is an increasing delay time and a decrease of the proportional band. Tests on the analogue computer have shown that instabilities can occur.



In spite of this the transient response of the controller arrangement Figure 4 shows an increase of the delay time corresponding to the error. The reset time  $T_n$  and the proportional band are constant. This can be proved by analytic methods. The damping ratio of the transient response increases.

As an example, in Figures C and D can be seen the transient responses of the closed loop systems for step changes of desired value. The controllers used for the test on the analogue computer both had

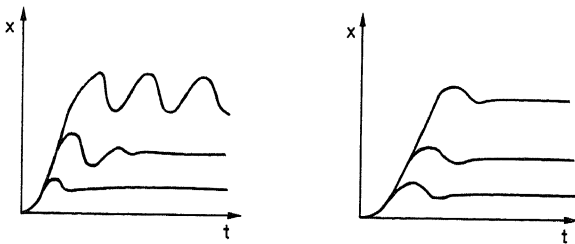


Figure C. Transient response for step changes of desired value (PI controller, Fig. 3)

Figure D. Transient response for step changes of desired value (PI controller, Fig. 4)

the same settings as the ideal linear PI controller, and the same running time of the motor. The result proves that the arrangement of Figure 4 is superior to that of Figure 3 as soon as the running time has an essential influence on the dynamic behaviour.

(2) Contrary to the author's opinion, there is no advantage in using relays. The results are the same if the relay (Figure 4) is replaced by a high gain amplifier with restricted output signal. The unfavourable influence of the sensitivity of the relay is not taken into consideration.

The advantages appear only in the feedback circuit. The following arrangement (Figure E) is equal in dynamic behaviour to Figure 4.

(3) As already pointed out in earlier publications<sup>1</sup> the controller Figure 4 is linear if the charge and discharge time constants of the feedback system are equal, if the restricted speed of the actuator is neglected. Unequal time constants result in an increasing or decreasing proportional band. Thus, in accordance with Dr. Pavlik, the quality of control decreases.

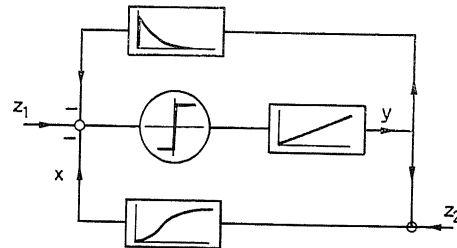


Figure E

(4) Dr. Pavlik points out that a closer approximation to the optimum control function will be achieved by adding a time lag to the controller or by increasing the running time of the motor. By this the damping ratio of the transient response increases. At the same time the integral of the absolute error increases too. In practice this effect is most undesirable.

#### Reference

- <sup>1</sup> BÖTTCHER, W. Vergleich von Dreipunktreglern mit einem linearen kontinuierlichen PI-Regler. *Regelungstechnik* 10 (1962) 114-119 and 210-213

E. PAVLIK, in reply

(1) Many thanks to Mr. Böttcher for his discussion remarks. I am glad to see that he has essentially made the same observations when examining the systems shown in Figures 3 and 4, i.e. that the systems show a strongly different behaviour, the system of Figure 4 being unambiguously superior to that of Figure 3. This is the more pleasant as Mr. Böttcher has taken the trouble to deal with the problems with a special modification of linear methods and not on the plane of the maximum principle. Generally valid determinations can hardly be made in this way, it is true, the systems being fundamentally non-linear in their character and can only be described by the exact methods of non-linear control theory.

(2) The maximum principle requires the most exact utilization of the limited coordinate. Consequently, I cannot agree with the opinion of Mr. Böttcher that the application of relays brings no advantages. It is true that the control behaviour of the system shown in Figure 4 is not substantially changed if the three-point switch is substituted by a continuous amplifier with a high gain degree and limited output signal, provided that the dynamic range of the amplifier is about identical with the response range (dead zone) of the relay. But in this case the continuous amplifier is substantially nothing but a relay.

(4) This remark is the result of the translation error already mentioned in my reply to the contribution of Mr. Schreiner. Mr. Böttcher is also right, of course, in that an artificial time lag of the servomechanism can only make the total control behaviour worse.

# Nouvelle Procédure d'Optimalisation Statistique Fondée sur la Transformation

$$V = (z - 1) / (z + 1)$$

P. LEFÈVRE

## Summary

This paper deals with the optimization of sampled data systems subjected to several constraints and operating on noisy inputs; while signal and noise are supposed to be a random stationary, ergodic and stochastically independent time series, the optimization criterion is the R.M.S. error and saturating signals minimization.

The derivation of the optimum system is greatly simplified by using the  $V \triangleq (z-1)/(z+1)$  transform, since the computation is done with the mathematical tools available for continuous systems.

The derivation of the optimum digital compensator in the  $V$ -plane suggests a new optimization procedure as an extension of Newton's theory.

This method is applied to the following problem. Given (a) a zero order hold cascaded with a plant whose transference is  $H_A(s) = 1/[s^2(1+\tau s)^2]$ ; (b) the aforementioned input spectra, and (c) two acceleration constraints, find the optimum digital compensator to be inserted in the loop.

The algebraic equations leading to the solution of this problem are thoroughly developed hereafter; numerical applications refer to the influence of the plant time constant  $\tau$  on the performance index; this is particularly important with digital systems, for it narrows the optimization zone; the greater this time constant is, the smaller the sampling period should be made.

This paper also includes a table of the most usual  $V$ -transforms of transfer functions  $F(s)$ .

## Sommaire

Le présent mémoire traite de l'optimalisation des systèmes échantillonnés soumis à plusieurs contraintes et opérant en présence de bruit; le message et le bruit étant aléatoires stationnaires, ergodiques et indépendants. Le critère retenu est la minimisation des valeurs quadratiques moyennes de l'erreur et des grandeurs saturantes.

La procédure d'optimalisation du système numérique se trouve très simplifiée par l'emploi de la transformée  $V^2$  où  $V \triangleq (z-1)/(z+1)$ ; tous les calculs s'effectuant alors avec l'outillage mathématique propre aux systèmes analogiques.

L'élaboration, dans le domaine ( $V$ ), du compensateur arithmétique optimal suggère une nouvelle méthode d'optimalisation qui constitue une extension de la théorie de Newton.

Cette méthode a été appliquée au problème ci-après: optimalisation, avec deux contraintes en accélération, d'un compensateur numérique inséré dans un système bouclé, comprenant un circuit bloqueur d'ordre zéro et un élément analogique:  $H_A(s) = 1/[s^2(1+\tau s)^2]$ .

Sont indiquées toutes les relations littérales nécessaires à la résolution de ce problème. Les applications numériques portent sur l'influence de la constante de temps  $\tau$ , à l'égard des performances optimales.

Celle-ci prend une importance particulière en numérique, car elle restreint l'intervalle d'optimalisation. Toute apparition ou élévation de constante de temps devra, le plus souvent, être compensée par un abaissement de la période d'échantillonnage.

Ce mémoire présente également une table de transformées en  $V$ , les plus courantes, concernant les fonctions  $F(s)$ .

## Zusammenfassung

Dieser Beitrag befaßt sich mit der Optimierung von Abtastsystemen bei rauschförmigem Eingangssignal, die verschiedenen Beschränkungen unterliegen. Da das Nutz- und das Rauschsignal als zufallsstationäre, ergodische und stochastisch unabhängige Zeitreihen angenommen werden, gelten als Optimierkriterien der Fehler des Effektivwertes und ein Minimum an hohen Signalspitzen. Die Ermittlung des optimalen Systems wird durch Verwendung der Transformation  $V \triangleq (z-1)/(z+1)$  wesentlich erleichtert, da sich die Rechnungen mit den für kontinuierliche Systeme verfügbaren mathematischen Mitteln durchführen lassen.

Die Suche nach dem optimalen digitalen Kompensator in der  $V$ -Ebene führt zu einem neuen Optimierverfahren, das eine Erweiterung der Newton'schen Theorie darstellt.

Die Methode wird auf das folgende Problem angewendet:

- a) ein Halteglied nullter Ordnung in Reihe mit einer Regelstrecke mit der Übertragungsfunktion  $H_A(s) = 1/[s^2(1+\tau s)^2]$ ,
  - b) die oben genannten Eingangsspektren und
  - c) zwei Beschränkungen für die Beschleunigung;
- gesucht wird der optimale in den Regelkreis einzuführende digitale Kompensator.

Danach werden die algebraischen Gleichungen, die zur Lösung dieses Problems führen, ausführlich abgeleitet; die numerische Berechnung dient vor allem der Untersuchung des Einflusses der Zeitkonstanten  $\tau$  der Regelstrecke auf die Gütezahl; dies ist bei digitalen Systemen besonders wichtig, weil der Optimierungsbereich dadurch eingengt wird. Je größer diese Zeitkonstante ist, um so kleiner soll im allgemeinen die Abtastperiode gewählt werden.

Der Beitrag enthält daneben eine Tabelle der gebräuchlichsten  $V$ -Transformationen der Übertragungsfunktionen  $F(s)$ .

## Introduction

L'optimalisation statistique du système échantillonné, sollicité par des signaux aléatoires stationnaires et ergodiques, a été traitée jusqu'à ce jour uniquement à l'aide de la transformée en  $z^{1-5}$ . Bien que l'emploi de cette transformée soit tout à fait justifié du point de vue théorique, il n'en complique pas moins l'analyse des cas concrets, pour les raisons énumérées ci-après.

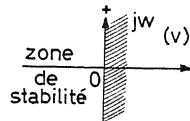
(1) L'optimalisation en  $z$  engendre des opérations de base très différentes — comme le montre le *Tableau 1* — de celles requises par l'optimalisation des systèmes continus, laquelle se développe dans le domaine de Laplace ( $s$ ). L'optimalisation du système échantillonné implique donc la mise au point de programmations spéciales.

(2) Comme nous l'ont révélé de nombreuses applications traitées à l'aide d'un ordinateur 650 IBM, l'optimalisation en  $z$  devient rapidement imprécise et conduit même parfois à des résultats illusoire, dès que se complique tant soit peu la structure de l'élément analogique fixé *a priori*. Cette perte de précision est une conséquence directe de l'élévation du nombre de pôles



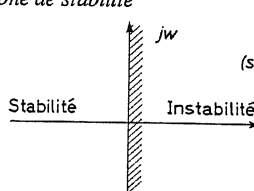
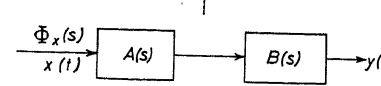
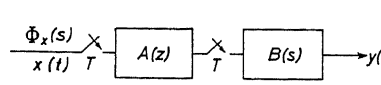
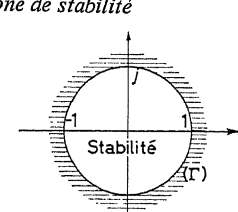
et de zéros stables — lesquels se trouvent alors resserrés les uns contre les autres à l'intérieur du cercle unité — et de la nécessité, d'évaluer, par la seule méthode des résidus de Cauchy, les intégrales circulaires du genre de celle figurant dans le Tableau 1.

Cette situation nous a incité à entreprendre l'optimisation du système échantillonné à l'aide de la transformée en  $v$ , déduite de la précédente par la relation

$$V \triangleq \frac{z-1}{z+1} \quad (1)$$


L'intérêt d'une telle opération est de transformer le cercle unité du domaine ( $z$ ) en l'axe imaginaire  $jw$  du domaine ( $v$ ) et

Tableau 1. Principales opérations effectuées au cours d'une optimisation statistique

Optimisation du système continu dans le domaine ( $s$ )	Optimisation du système échantillonné dans le domaine ( $z$ )
<p><b>Zône de stabilité</b></p>  <p><b>Décomposition d'une fonction paire par rapport à l'axe imaginaire <math>jw</math></b></p> $F(s) = f(s) \cdot f(-s)$ <p>où <math>f(s)</math> renferme tous les pôles et zéros stables de <math>F(s)</math> lesquels sont situés dans le demi-plan gauche.</p> <p><b>Evaluation de la valeur quadratique moyenne <math>\overline{y^2(t)}</math> en fonction de la densité spectrale <math>\Phi_y(s)</math></b></p> <p>* au-dessous</p> $\overline{y^2(t)} = \frac{1}{2\pi j} \int_{-j\infty}^{j\infty} \Phi_x(s) A(s) \cdot A(-s) \cdot B(s) \cdot B(-s) \cdot ds$ <p>et puisque <math>\Phi_x(s)</math> est une fonction paire</p> $\overline{y^2(t)} = \frac{1}{\pi} \int_0^\infty \varphi_x(jw) A(jw) B(jw)^2 dw$ <p>avec</p> $\Phi_x(s) = \varphi_x(s) \cdot \varphi_x(-s)$ <p>* </p> <p>† </p>	<p><b>Zône de stabilité</b></p>  <p><b>Décomposition d'une fonction réciproque par rapport au cercle unité</b></p> $F(z) = f(z) f(z^{-1})$ <p>où <math>f(z)</math> renferme tous les pôles et zéros stables de <math>F(z)</math> lesquels sont situés à l'intérieur du cercle unité.</p> <p><b>Evaluation de la valeur quadratique moyenne à partir de la densité spectrale échantillonnée</b></p> <p>† au-dessous</p> $\overline{y^2(t)} = \frac{1}{2\pi j} \oint_{\text{cercle unité}} A(z) \cdot A(z^{-1}) \cdot \Phi_x(z) Z[B(s) \cdot B(-s)] \frac{dz}{Tz}$ <p>la transposition dans le domaine (<math>v</math>) s'effectue aisément. On trouve:</p> $\overline{y^2(t)} = \frac{1}{T\pi j} \int_{-j\infty}^{j\infty} A(v) A(-v) \Phi_x(v) V[B(s) B(-s)] \frac{dv}{1-v^2} \quad (5)$ <p>Les fonctions <math>\Phi_x(v)</math> et <math>V[B(s) B(-s)]</math>, étant paires, leur décomposition par rapport à l'axe imaginaire <math>jw</math> s'effectue comme celle de la fonction <math>F(s)</math> du Tableau I:</p> $\begin{cases} \Phi_x(v) = \varphi_x(v) \varphi_x(-v) \\ V[B(s) B(-s)] = b(v) b(-v) \end{cases} \quad (6)$

d'étendre ainsi la zone de stabilité de ce dernier domaine à la totalité du demi-plan de gauche. L'analogie existant entre les zones de stabilité respectives des domaines ( $v$ ) et ( $s$ ) simplifie alors beaucoup la procédure d'optimisation du système échantillonné en faisant bénéficier celle-ci de tout l'outillage mathématique (Tables<sup>7,8</sup>, Programmes etc. ...) actuellement disponible pour l'optimisation du système analogique.

Ce mémoire débutera par la présentation d'une table permettant de passer directement d'une fonction  $F(s)$  à sa transformée en  $v$ :  $V[F(s)]$  ou  $F(v)$  — puis exposera, dans ses grandes lignes, la méthode d'optimisation statistique en ( $v$ ) applicable au système échantillonné, en présence de bruit et de plusieurs contraintes. Il s'achèvera sur un exemple d'application, examinant l'influence exercée par les constantes de temps de l'élément imposé, à l'égard du choix de la période de prélèvement  $T$ .

### Généralités sur la transformée en $V$

#### Définition et mode d'utilisation

Soit une fonction  $f(t)$  définie pour tout  $t$  et possédant une transformée de Laplace

$$F(s) = \int_{-\infty}^{\infty} f(t) e^{-st} dt$$

la transformée en  $z$  bilatérale est définie par

$$F(z) \triangleq \sum_{n=-\infty}^{\infty} f(nT) z^{-n} \quad (2)$$

et la transformée en  $v$  qui lui est associée s'écrit:

$$F(v) \triangleq [F(z)]_{z=\frac{1+v}{1-v}} = \sum_{n=-\infty}^{\infty} f(nT) \left( \frac{1-v}{1+v} \right)^n \quad (3)$$

La table, reportée à l'annexe I, indique les transformées en  $v$  des fonctions  $F(s)$  les plus couramment rencontrées dans un problème d'optimisation.

En ce qui concerne son mode d'utilisation, la transformée en  $v$  obéit strictement aux mêmes règles que la transformée en  $z$ .

#### Expression de la valeur quadratique moyenne dans le domaine $v$

La méthode de Sklansky<sup>6</sup>, appliquée au système échantillonné du Tableau I, permet d'écrire la valeur quadratique moyenne de la sortie  $y(t)$  sous la forme suivante:

$$\overline{y^2(t)} = \frac{1}{2\pi j} \oint_{\text{cercle unité}} A(z) A(z^{-1}) \Phi_x(z) Z[B(s) B(-s)] \frac{dz}{Tz} \quad (4)$$

la transposition dans le domaine ( $v$ ) s'effectue aisément. On trouve:

$$\overline{y^2(t)} = \frac{1}{T\pi j} \int_{-j\infty}^{j\infty} A(v) A(-v) \Phi_x(v) V[B(s) B(-s)] \frac{dv}{1-v^2} \quad (5)$$

Les fonctions  $\Phi_x(v)$  et  $V[B(s) B(-s)]$ , étant paires, leur décomposition par rapport à l'axe imaginaire  $jw$  s'effectue comme celle de la fonction  $F(s)$  du Tableau I:

$$\begin{cases} \Phi_x(v) = \varphi_x(v) \varphi_x(-v) \\ V[B(s) B(-s)] = b(v) b(-v) \end{cases} \quad (6)$$



Dans ces conditions, la valeur quadratique moyenne (5) s'exprime comme suit :

$$\overline{y^2(t)} = \frac{2}{T\pi} \int_0^\infty \left| \frac{A(j\omega) \varphi_x(j\omega) b(j\omega)}{1+j\omega} \right|^2 d\omega \quad (7)$$

revêtant ainsi la même forme que pour le système continu.

### Procédure d'optimalisation dans le domaine ( $V$ )

La méthode, exposée dans ses grandes lignes ci-dessous, constitue une extension aux systèmes échantillonnés, des théories d'optimalisation de Wiener et Newton<sup>7</sup> élaborées pour les systèmes continus. Elle s'apparente donc à la procédure utilisée par divers auteurs, notamment par Chang<sup>2</sup>, Tou<sup>3</sup>, Thellier<sup>5</sup>, au cours de leurs travaux dans le domaine ( $z$ ). Elle va être précisée à l'aide du système bouclé de la Figure 1. Celui-ci est supposé recevoir

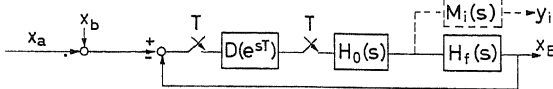


Figure 1. Schéma de l'asservissement considéré

deux signaux aléatoires stationnaires, ergodiques et non corrélés : le message  $x_a(t)$  et le bruit  $x_b(t)$ , dont les densités spectrales respectives  $\Phi_a(s)$  et  $\Phi_b(s)$  sont connues.

Ce système renferme, d'une part, un circuit bloqueur d'ordre zéro  $H_0(s)$  et des éléments analogiques, linéaires et invariants dans le temps, dont les transmittances  $H_f(s)$  et  $M_i(s)$  sont fixées *a priori*, et, d'autre part, un compensateur arithmétique, de transmittance  $D(e^{sT})$  inconnue, sur lequel repose l'optimalisation.

Le problème consiste à élaborer ce compensateur de façon à rendre minimales les valeurs quadratiques moyennes de l'erreur :  $\varepsilon(s) \triangleq x_a(s) - x_E(s)$  et de certaines grandeurs saturantes  $y_i(s)$ .

Cela revient à prendre, comme critère d'optimalisation, la minimalisation de la quantité suivante :

$$Q_n = \overline{\varepsilon^2(t)} + \sum_{i=1}^n \lambda_i \overline{y_i^2(t)} \quad (8)$$

où,  $\lambda_i$ , désigne le paramètre de Lagrange associé à la contrainte relative au signal saturant  $y_i$ .

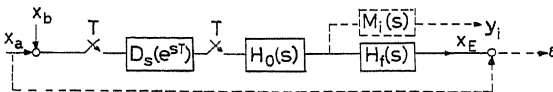


Figure 2. Représentation fonctionnelle de l'optimalisation statistique

On passe aisément du système bouclé initial au montage en série équivalent de la Figure 2, en posant :

$$D_s(v) = \frac{D(v)}{1 + D(v)Q(v)} \text{ avec } Q(s) \triangleq H_0(s)H_f(s) \quad (9)$$

Les valeurs quadratiques moyennes s'expriment comme suit :

$$\begin{aligned} \overline{\varepsilon^2(t)} &= \overline{x_a^2(t)} \\ &+ \frac{1}{\pi T j} \int_{-j\infty}^{j\infty} \{ -D_s(v) V [Q(s) \Phi_a(s)] \\ &- D_s(-v) V [Q(-s) \Phi_a(-s)] \\ &+ D_s(v) D_s(-v) [\Phi_a(v) + \Phi_b(v)] \cdot V [Q(s) Q(-s)] \} \frac{dv}{1-v^2} \end{aligned} \quad (10)$$

$$- D_s(-v) V [Q(-s) \Phi_a(-s)]$$

$$+ D_s(v) D_s(-v) [\Phi_a(v) + \Phi_b(v)] \cdot V [Q(s) Q(-s)] \} \frac{dv}{1-v^2} \quad (10)$$

$$\overline{y_i^2(t)} = \frac{1}{\pi T j} \int_{-j\infty}^{j\infty} D_s(v) D_s(-v) (\Phi_a(v) + \Phi_b(v))$$

$$\cdot V [M_i(s) M_i(-s) H_0(s) H_0(-s)] \frac{dv}{1-v^2} \quad (11)$$

et la quantité  $Q_n$  revêt alors la forme suivante :

$$\begin{aligned} Q_n &= \overline{x_a^2(t)} + \frac{1}{T\pi j} \int_{-j\infty}^{j\infty} \\ &[-D_s(v) F(v) - D_s(-v) F(-v) + D_s(v) D_s(-v) C(v)] \frac{dv}{1-v^2} \end{aligned} \quad (12)$$

avec :

$$\begin{cases} F(-v) \triangleq V [Q(-s) \Phi_a(s)] \\ C(v) \triangleq \left\{ V [Q(s) Q(-s)] \right. \\ \quad \left. + \sum_{i=1}^n \lambda_i V [M_i(s) M_i(-s) H_0(s) H_0(-s)] \right\} \\ \quad \cdot (\Phi_a(v) + \Phi_b(v)) \end{cases} \quad (13)$$

Le compensateur optimal, qui sera désigné par,  $D_{s0}(v)$ , devant rendre minimale la quantité  $Q_n$ , se déduit de la relation (12) au moyen du calcul des variations. Cela consiste à poser  $D_s(v) = D_{s0}(v) + \varrho \delta(v)$  où,  $\varrho$ , est un paramètre réel et,  $\delta(v)$ , une fonction appartenant au même ensemble que la fonction  $D_{s0}(v)$  et à satisfaire la condition

$$\left( \frac{\partial Q_n}{\partial \varrho} \right)_{\varrho=0} = 0$$

La minimalisation de  $Q_n$  est obtenue pour :

$$\int_{-j\infty}^{j\infty} \delta(-v) \left[ \frac{-F(-v) + D_{s0}(v) C(v)}{1-v^2} \right] dv = 0$$

Cette équation implique l'analyticité, dans le demi-plan gauche, de la quantité suivante :

$$-\frac{F(-v)}{(1+v)C^-(v)} + \frac{D_{s0}(v)C^+(v)}{(1+v)}$$

en posant :

$$C(v) \triangleq C^+(v) \cdot C^-(v), \quad (14)$$

$C^+(v)$  contenant tous les pôles et zéros stables de  $C(v)$ .

Il en résulte l'expression du compensateur équivalent optimal :

$$\begin{aligned} D_{s0}(v) &= \frac{(1+v)}{C^+(v)} \left[ \frac{F(-v)}{(1+v)C^-(v)} \right]_+ \\ &\quad \left[ \frac{F(-v)}{(1+v)C^-(v)} \right]_+ \end{aligned} \quad (15)$$

désignant la partie de la décomposition de la fonction rationnelle

$$\frac{F(-v)}{(1+v)C^-(v)}$$

en éléments simples qui contient tous les pôles stables.

L'expression du compensateur optimal  $D_0(v)$ , à insérer dans le système bouclé initial, découle de (9) et (15).

### Exemple d'application de l'optimisation en $V$

#### Présentation du problème

On se propose d'optimiser le système schématisé à la Figure 3, dans lequel la transmittance imposée renferme une double constante de temps,  $\tau$ , montée en série avec une double intégration:

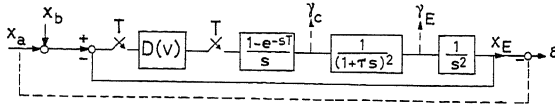


Figure 3. Structure du système à optimiser

$$H_f(s) \triangleq \frac{1}{(1+\tau s)^2} \times \frac{1}{s^2} \quad (16)$$

les signaux d'entrée font l'objet des données suivantes:  
— la densité spectrale du message  $x_a$  est définie par:

$$\Phi_a(s) \triangleq \frac{a^2}{s^4(b^2 - s^2)} \quad (17)$$

— le bruit  $x_b$  est supposé blanc dans l'intervalle

$$\left(-\frac{\pi}{T} \leq \omega \leq \frac{\pi}{T}\right)$$

de niveau  $k^2$  et nul ailleurs.

Le problème consistera d'abord à rechercher le compensateur arithmétique optimal  $D_0(v)$  qui minimise la quantité

$$Q_2 = \overline{\varepsilon^2(t)} + \lambda_C \cdot \overline{\gamma_C^2(t)} + \lambda_E \cdot \overline{\gamma_E^2(t)} \quad (18)$$

dans laquelle interviennent les paramètres de Lagrange,  $\lambda_C$  et  $\lambda_E$ , ainsi que les valeurs quadratiques moyennes de l'erreur  $\varepsilon$ , de l'accélération de sortie  $\gamma_E$  et de l'accélération commandée  $\gamma_C$ .

Il s'agira ensuite d'évaluer les performances réalisées par l'asservissement optimal donc de chiffrer, pour tout ensemble  $(\lambda_C, \lambda_E)$  fixé, les grandeurs suivantes:

(a) Erreur efficace:

$$\sigma_\varepsilon \triangleq (\overline{\varepsilon^2(t)})^{\frac{1}{2}}$$

(b) Accélération de sortie efficace:

$$\sigma_{\gamma_E} \triangleq (\overline{\gamma_E^2(t)})^{\frac{1}{2}}$$

(c) Accélération commandée efficace:

$$\sigma_{\gamma_C} \triangleq (\overline{\gamma_C^2(t)})^{\frac{1}{2}}$$

#### Calculs d'optimisation

Les diverses expressions, engendrées par l'application de la théorie que nous avons développée précédemment, vont être précisées à l'aide des deux paramètres:

$$x \triangleq \frac{T}{\tau} \quad (19)$$

$$u \triangleq bT \quad (20)$$

et des coefficients  $A_i$ ,  $\Phi_i$ ,  $L_i$ ,  $M_i$ ,  $G_i$  et  $r_i$  explicités à l'Appendice II.

Calcul des transformées en  $v$  —

$$\Phi_a(v) = a^2 V \left[ \frac{1}{s^4(b^2 - s^2)} \right]; \Phi_b(v) = \frac{k^2}{T}$$

$$\Phi_a(v) + \Phi_b(v) = \frac{a^2 T^5}{16 u^2} \cdot \frac{A(v)}{v^4 \left( 1 - \coth^2 \frac{u}{2} v^2 \right)} \quad (21)$$

en posant:

$$A(v) \triangleq 1 + A_2 v^2 + A_4 v^4 + A_6 v^6 \quad (22)$$

Calcul de  $C(v)$ . D'après (13), on a

$$C(v) = K(v) [\Phi_a(v) + \Phi_b(v)] \quad (23)$$

en posant

$$K(v) = V \left[ \frac{(1 - e^{-Ts})(1 - e^{Ts})}{-s^6(1 - \tau^2 s^2)^2} \right] + \lambda_C T + \lambda_E V \left[ \frac{(1 - e^{-Ts})(1 - e^{Ts})}{-s^2(1 - \tau^2 s^2)^2} \right] \quad (24)$$

Cette dernière fonction dépend donc à la fois de la structure de l'élément analogique imposé et des contraintes retenues. En revanche, elle reste indépendante des signaux d'entrée. Les calculs procurent:

$$K(v) = \frac{T^5}{16} \cdot \frac{\Phi(v)}{v^4 \left[ 1 - \coth^2 \frac{x}{2} v^2 \right]^2} \quad (25)$$

avec

$$\Phi(v) \triangleq 1 + \Phi_2 v^2 + \Phi_4 v^4 + \Phi_6 v^6 + \Phi_8 v^8 \quad (26)$$

Les relations (21), (23) et (25) conduisent à:

$$C(v) = \frac{a^2 T^{10}}{(16)^2 u^2} \cdot \frac{A(v) \Phi(v)}{v^8 \left[ 1 - \coth^2 \frac{u}{2} v^2 \right] \left[ 1 - \coth^2 \frac{x}{2} v^2 \right]^2} \quad (27)$$

Calcul de  $F(-v)$ .

$$F(-v) = -a^2 V \left[ \frac{1 - e^{Ts}}{s^7(1 - \tau s)^2(b^2 - s^2)} \right] \quad (28)$$

le développement d'une telle transformée ne peut être indiqué ici, en raison du manque de place.

#### Expression des compensateurs arithmétiques

Compensateur série équivalent optimal. Le développement de (28) joint à la relation (27) permet d'écrire la relation (15) sous la forme suivante:

$$D_{s0}(v) = \frac{4}{T^2} \cdot \frac{v^2(1+v) \left( 1 + \coth \frac{x}{2} v \right)^2 L(v)}{\alpha(v) \varphi(v)} \quad (29)$$

en posant:  $L(v) \triangleq 1 + L_1 v + L_2 v^2$  (30)

et en désignant respectivement par  $\alpha(v)$  et  $\varphi(v)$  les polynômes résultant de la décomposition, par rapport à l'axe imaginaire, des fonctions  $A(v)$  et  $\Phi(v)$ :

$$\begin{cases} \alpha(v) \alpha(-v) = A(v) \\ \varphi(v) \varphi(-v) = \Phi(v) \end{cases} \quad (31)$$

Ils se développent comme suit:

$$\begin{cases} \alpha(v) = 1 + \alpha_1 v + \alpha_2 v^2 + \alpha_3 v^3 \\ \phi(v) = 1 + \phi_1 v + \phi_2 v^2 + \phi_3 v^3 + \phi_4 v^4 \end{cases} \quad (32)$$

*Compensateur optimal  $D_0(v)$ .* Il se déduit de l'expression de  $D_{s0}(v)$  et de la relation (9). Cette dernière renferme la transformée  $Q(v)$ , définie par:

$$Q(v) = V \left[ \frac{1 - e^{-sT}}{s^3 (1 + \tau s)^2} \right] = \frac{T^2 (1 - v) M(v)}{4 v^2 \left( 1 + \coth \frac{x}{2} v \right)^2} \quad (33)$$

avec:

$$M(v) \triangleq 1 + M_1 v + M_2 v^2 + M_3 v^3 \quad (34)$$

On obtient ainsi l'expression suivante:

$$D_0(v) = \frac{4}{G_0 T^2} \frac{(1 + v) \left( 1 + \coth \frac{x}{2} v \right)^2 \cdot L(v)}{g(v)} \quad (35)$$

ou

$$g(v) \triangleq 1 + g_1 v + g_2 v^2 + g_3 v^3 + g_4 v^4 + g_5 v^5 \quad (36)$$

avec  $g_i \triangleq G_i / G_0$ , les coefficients  $G_i$  étant définis à l'Appendice II.

*Expression des valeurs quadratiques moyennes* — Avec les fonctions introduites plus haut, les valeurs quadratiques prennent les formes simples suivantes:

$$\text{Erreur} \quad \overline{\varepsilon^2(t)} \triangleq \overline{\varepsilon_a^2(t)} + \overline{\varepsilon_b^2(t)}$$

*Composante message:*

$$\overline{\varepsilon_a^2(t)} = \frac{a^2 T^5 G_0^2}{8 \pi u^2} \int_0^\infty |E_a(j\omega)|^2 d\omega$$

avec

$$E_a(v) \triangleq \frac{r(v) g(v)}{\left( 1 + \coth \frac{u}{2} v \right) \alpha(v) \varphi(v)}$$

et

$$r(v) = 1 + r_1 v + r_2 v^2$$

*Composante bruit:*

$$\overline{\varepsilon_b^2(t)} = \frac{2 k^2}{\pi T} \int_0^\infty |E_b(j\omega)|^2 d\omega$$

avec

$$E_b(v) \triangleq \frac{(1 + v) L(v) M(v)}{\alpha(v) \varphi(v)}$$

*Accélération de sortie:*

$$\overline{\gamma_E^2(t)} = \frac{2 a^2 T}{\pi u^2} \int_0^\infty |\Gamma_E(j\omega)|^2 d\omega \quad (37)$$

avec

$$\Gamma_E(v) \triangleq \frac{(1 + v) \left[ 1 + \left( \coth \frac{x}{2} - \frac{x}{2} \left( \coth^2 \frac{x}{2} - 1 \right) \right) v \right] \cdot L(v)}{\left( 1 + \coth \frac{u}{2} v \right) \varphi(v)}$$

*Accélération commandée:*

$$\overline{\gamma_C^2(t)} = \frac{2 a^2 T}{\pi u^2} \int_0^\infty |\Gamma_C(j\omega)|^2 d\omega \quad (38)$$

avec

$$\Gamma_C(v) \triangleq \frac{\left( 1 + \coth \frac{x}{2} v \right)^2 L(v)}{\left( 1 + \coth \frac{u}{2} v \right) \varphi(v)}$$

*Applications numériques: influence de la constante de temps  $\tau$  sur le choix de la période de prélèvement*

Cette étude a nécessité la résolution de 40 équations de 6e ou 8e degré et le calcul de 200 intégrales environ.

*Données numériques* — Les résultats présentés ci-après concernent les données suivantes:  $a^2 = 9 \text{ (m}^2/\text{sec}^5)$ ;  $k^2 = 14 \text{ (m}^2/\text{rad/sec)}$ ;  $b = 2/45 \text{ (sec}^{-1})$  et  $\lambda_E = 0$ . L'optimisation s'effectue donc avec la seule contrainte en accélération commandée:  $\lambda_c \neq 0$ . Les paramètres de l'étude sont constitués par la période  $T$  et la constante de temps  $\tau$  de l'élément analogique.

*Présentation des résultats*

*Courbes de performances et point d'arrêt.* On désigne par courbe de performances, le graphique de la variation de l'erreur efficace  $\sigma_e$  en fonction de la valeur efficace de l'accélération de sortie  $\sigma_{\gamma_E}$ , relatif à une valeur fixée des paramètres  $T$  et  $\tau$ .

La Figure 4 (a) montre deux courbes de ce type relatives à  $\tau = 0,1 \text{ sec}$  et à 2 valeurs de  $T$ . Le cas  $T = 0$  représente le cas analogique traité selon la méthode de Newton. On constate que la courbe  $T = 0$  est asymptote à une certaine valeur  $(\sigma_e)_L$  correspondant à l'optimisation de Wiener ( $\lambda_E = 0$ ), tandis que la courbe  $T = 0,4 \text{ sec}$ , toujours située au-dessus de la courbe précédente, présente un point d'arrêt d'abscisse  $(\sigma_{\gamma_E})_M$ . L'existence d'un tel point s'explique mathématiquement du fait que le degré de la fonction  $\Phi(v)$  reste invariant lorsque s'annulent les paramètres  $\lambda_E$  et  $\lambda_c$  [voir relation (37) et Annexe II].

On sait déjà<sup>5</sup> que l'abscisse  $(\sigma_{\gamma_E})_M$  du point d'arrêt décroît, lorsque  $T$  augmente, et qu'elle constitue ainsi un élément fondamental pour le choix de cette période.

La Figure 4 (b) indique les variations de l'accélération commandée,  $\sigma_{\gamma_C}$ , le long des courbes précitées. Le filtrage de cette accélération est plus énergétique en digital qu'en analogique.

*Influence de la constante de temps  $\tau$ .* Les points d'arrêt du système digital ont été calculés pour de nombreuses valeurs de  $T$  et de  $\tau$ . Ils font l'objet de la Figure 5 sur laquelle ont été tracés, en traits interrompus, les lieux de ces points à  $\tau$  constant, et, en traits pleins, les courbes de performances du système analogique concernant les mêmes valeurs de  $\tau$ . On constate que la constante

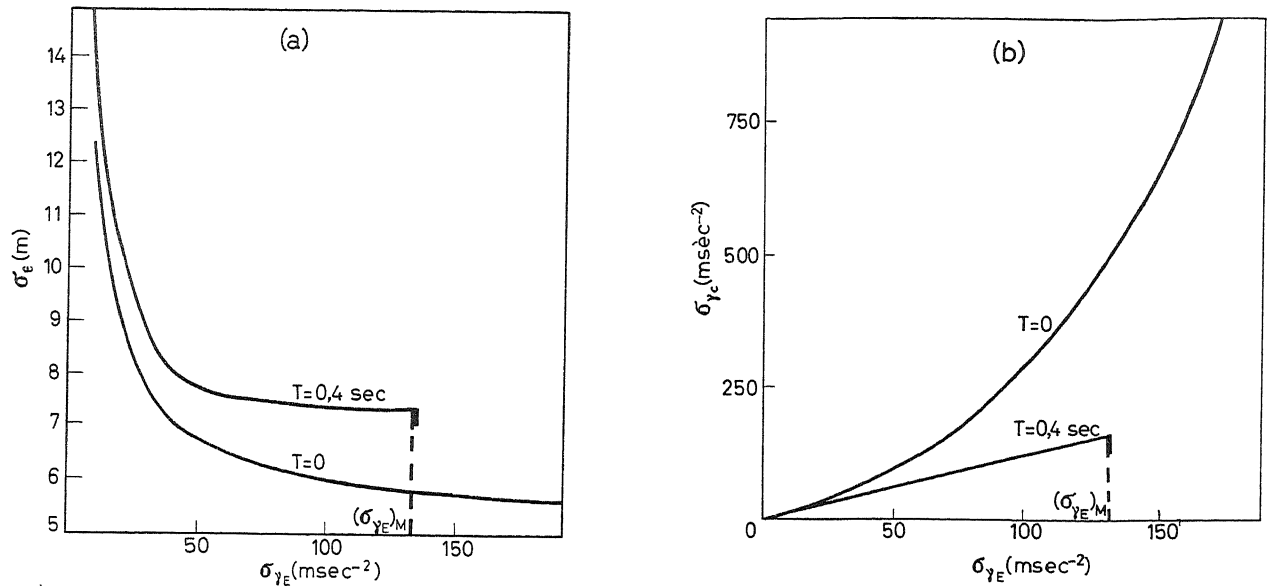


Figure 4. Courbes relatives à  $\tau = 0,1$  sec. (a) courbes de performances; (b) variations de l'accélération commandée

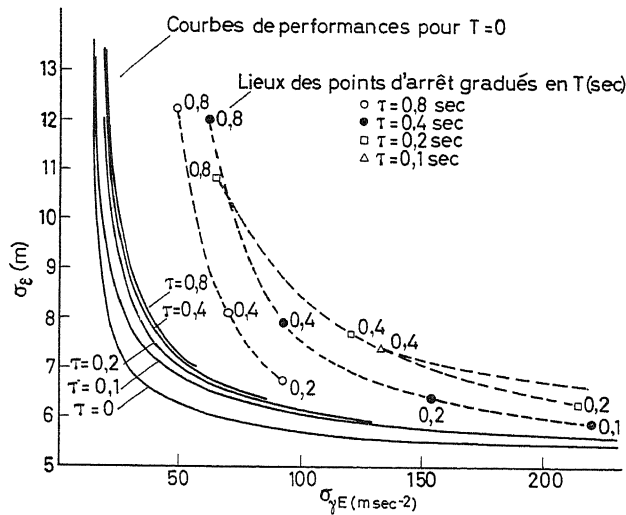


Figure 5. Lieux des points d'arrêt des courbes de performances à  $\tau$  constant

de temps n'intervient pratiquement pas dans l'accroissement de l'erreur consécutif au passage de l'analogique au digital; celui-ci étant essentiellement imputable à la période  $T$ . L'évaluation de cet accroissement, à la hauteur du point d'arrêt, procure:

$T$ (sec)	0,2	0,4	0,8
$\Delta\sigma_e$ (en %)	10	28	70

En revanche, la constante de temps joue un rôle très important à l'égard de l'intervalle d'optimisation, dont la borne supérieure est constituée par l'abscisse du point d'arrêt. Comme le montre la Figure 6, un accroissement de  $\tau$  se traduit toujours par une réduction de cette abscisse, d'autant plus accentuée que la période  $T$  est plus faible. Le maintien de l'intervalle d'optimisation sur une longueur donnée exige que tout accroissement de la constante de temps soit compensé par une réduction de la période  $T$ . Par exemple, le maintien de  $(\sigma_{\gamma E})_M$  sur 240 (m sec<sup>-2</sup>)

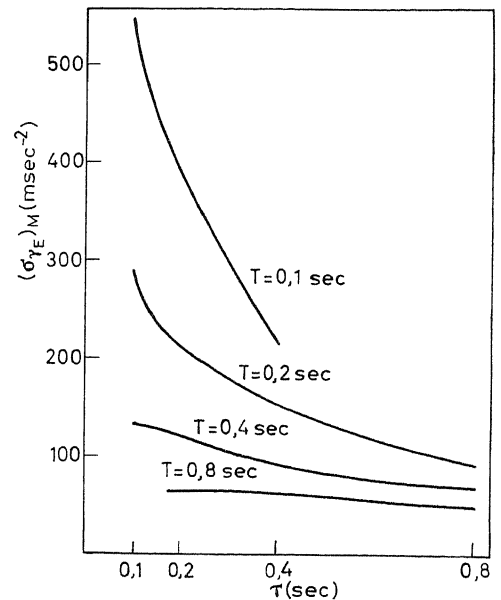


Figure 6. Variations de l'abscisse du point d'arrêt, avec  $\tau$

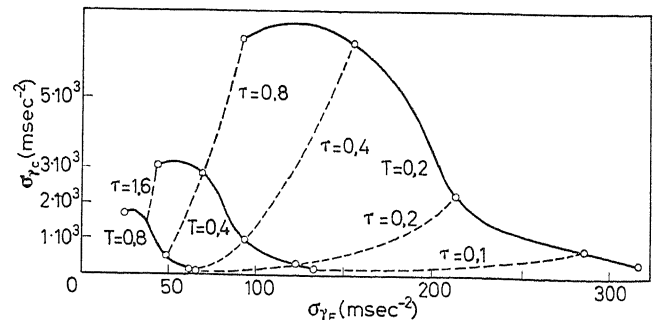
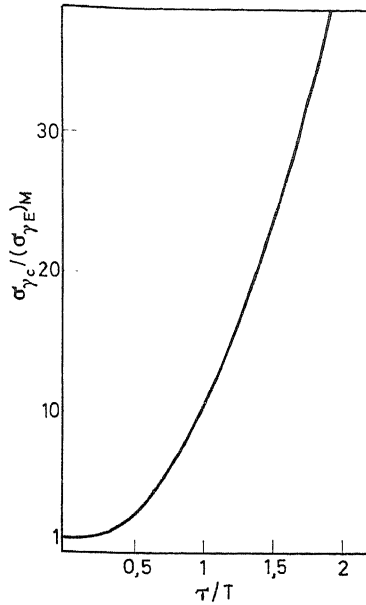


Figure 7. Accélération commandée pour les points d'arrêt de la Figure 5


 Figure 8. Variations de  $\sigma_{\gamma_e}/\sigma_{\gamma_E}$  déduites de la Figure 7

implique, lorsque  $\tau$  s'élève de 0,15 à 0,37 sec, que la période  $T$  soit réduite de 0,2 à 0,1 sec. Par ailleurs, l'accélération commandée efficace a été calculée en chacun des points d'arrêt de la Figure 5. Le graphique de la Figure 7 indique les variations de  $\sigma_{\gamma_e}$  en fonction de  $(\sigma_{\gamma_E})_M$ . L'exploitation de ce graphique conduit au tracé de la courbe unique de la Figure 8, d'après laquelle, le filtrage de  $\gamma_E$  par rapport à  $\gamma_e$  ne dépend que du quotient  $\tau/T$  et croît rapidement avec ce dernier. Un tel résultat pouvait, dans une certaine mesure, être prédit à l'aide des relations (37) et (38).

#### Conclusion

Les constantes de temps contenues dans l'élément imposé influent beaucoup plus sur l'optimalisation digitale que sur l'optimalisation analogique. Leur action sur celle-là est particulièrement néfaste puisque restreignant l'intervalle d'optimalisation. Toute apparition ou élévation de constante de temps doit, le plus souvent, être compensée par un abaissement de la période de prélèvement.

L'auteur adresse ses remerciements à la Direction des Etudes et Fabrications d'Armement grâce à laquelle cette recherche a pu être entreprise et menée à bonne fin.

#### Appendice 1

 Tableau des Transformées en  $v$ 

Fonctions $S$	Transformées en $v$	Fonctions $S$	Transformées en $v$
$\coth \frac{ST}{2}$	$v$	$\frac{1}{S}$	$\frac{1+v}{2v}$
$e^{ST}$	$\frac{1+v}{1-v}$	$\frac{1}{S^2}$	$\frac{T(1-v^2)}{2^2 v^2}$
$\frac{1}{S+a}$	$\frac{(1+v)\left(\coth \frac{x}{2} + 1\right)}{2\left(1 + \coth \frac{x}{2} v\right)}$	$\frac{1}{S^3}$	$\frac{T^2(1-v^2)}{2^3 v^3}$
$\frac{1}{(S+a)^2}$	$\frac{T(1-v^2)\left(\coth^2 \frac{x}{2} - 1\right)}{2^2 \cdot \left(1 + \coth \frac{x}{2} v\right)^2}$	$\frac{1}{S^4}$	$\frac{T^3(1-v^2)}{2^4 v^4} \left[1 - \frac{1}{3} v^2\right]$
$\frac{1}{(S+a)^3}$	$\frac{T^2(1-v^2)\left(\coth^2 \frac{x}{2} - 1\right)\left(\coth \frac{x}{2} + v\right)}{2^3 \cdot \left(1 + \coth \frac{x}{2} v\right)^3}$	$\frac{1}{S^5}$	$\frac{T^4(1-v^2)}{2^5 v^5} \left[1 - \frac{2}{3} v^2\right]$
$\frac{1}{(S+a)^4}$	$\frac{T^3(1-v^2)\left(\coth^2 \frac{x}{2} - 1\right)\left[-\frac{1}{3}\right]}{2^4 \cdot \left(1 + \coth \frac{x}{2} v\right)^4}$	$\frac{1}{S^6}$	$\frac{T^5(1-v^2)}{2^6 v^6} \left[1 - v^2 + \frac{2}{15} v^4\right]$
	$+ \frac{4}{3} v + v^2 + \left(1 - \frac{v^2}{3}\right) \coth^2 \frac{x}{2}$	$\frac{1}{S^7}$	$\frac{T^6(1-v^2)}{2^7 v^7} \left[1 - \frac{4}{3} v^2 + \frac{17}{45} v^4\right]$
		$\frac{1}{S^8}$	$\frac{T^7(1-v^2)}{2^8 v^8} \left[1 - \frac{5}{3} v^2 + \frac{11}{15} v^4 - \frac{17}{315} v^6\right]$

## Appendice I—continued

Fonctions $S$	Transformées en $v$	Fonctions $S$	Transformées en $v$
$\frac{1}{S^2 - a^2}$	$-\frac{1}{2a} \times \frac{(1-v^2) \coth \frac{x}{2}}{\left(1 - \coth^2 \frac{x}{2} v^2\right)}$	$\frac{1}{S^9}$	$\frac{T^8(1-v^2)}{2^9 v^9} \left[1 - 2v^2 + \frac{6}{5}v^4 - \frac{62}{315}v^6\right]$
<i>Nota</i>	On a posé $x \triangleq aT$	$\frac{1}{S^{10}}$	$\frac{T^9(1-v^2)}{2^{10} v^{10}} \left[1 - \frac{7}{3}v^2 + \frac{16}{9}v^4 - \frac{88}{189}v^6 + \frac{62}{2835}v^8\right]$

Appendice II—Détermination des coefficients relatifs à l'exemple d'application de l'optimisation en  $V$ Coefficients  $A_i$ 

$$A_2 = -\frac{4}{3} - \coth^2 \frac{u}{2} + \frac{4}{u^2}$$

$$A_4 = \frac{1}{3} - \frac{4}{u^2} + \frac{8}{u^3} \coth \frac{u}{2} + 4 \left( \frac{1}{3} - \frac{1}{u^2} \right) \coth^2 \frac{u}{2} + \frac{16k^2 u^2}{a^2 T^6}$$

$$A_6 = -\coth^2 \frac{u}{2} \cdot \left( \frac{1}{3} - \frac{4}{u^2} + \frac{8}{u^3} \coth \frac{u}{2} + \frac{16k^2 u^2}{a^2 T^6} \right)$$

$$\begin{aligned} \Phi_6 = & -\frac{8}{x^4} - 4 \left( \frac{1}{15} - \frac{4}{3x^2} + \frac{22}{x^4} \right) \coth^2 \frac{x}{2} + \left( \frac{8}{x^2} - 1 \right) \coth^4 \frac{x}{2} \\ & + \frac{112}{x^5} \coth \frac{x}{2} \\ & + \frac{16}{T^4} \left[ \frac{\lambda_E}{2} \left( -1 - 3 \coth^2 \frac{x}{2} + \frac{6}{x} \coth \frac{x}{2} \right) - 2 \lambda_C \coth^2 \frac{x}{2} \right] \end{aligned}$$

$$\begin{aligned} \Phi_8 = & \coth^2 \frac{x}{2} \left\{ -\frac{8}{x^4} - \frac{112}{x^5} \coth \frac{x}{2} + 2 \left( \frac{1}{15} - \frac{4}{3x^2} + \frac{28}{x^4} \right) \coth^2 \frac{x}{2} \right. \\ & \left. + \frac{16}{T^4} \left[ \frac{\lambda_E}{2} \left( -1 + 3 \coth^2 \frac{x}{2} - \frac{6}{x} \coth \frac{x}{2} \right) + \lambda_C \coth^2 \frac{x}{2} \right] \right\} \end{aligned}$$

Coefficients  $L_i$ 

$$L_1 = \alpha_1 + \varphi_1 + \frac{4}{x} - 2 \coth \frac{x}{2}$$

$$L_2 = \left( L_1 - \coth \frac{u}{2} \right) \coth \frac{u}{2}$$

$$+ \frac{64 \left( \coth \frac{u}{2} \right)^4 \left( 1 + \coth \frac{x}{2} \coth \frac{u}{2} \right)^2}{u^6 \left( 1 + \frac{u}{x} \right)^2 \alpha \left( \coth \frac{u}{2} \right) \varphi \left( \coth \frac{u}{2} \right)}$$

Coefficients  $\Phi_i$ 

$$\Phi_2 = -1 - 2 \coth^2 \frac{x}{2} + \frac{8}{x^2}$$

$$\begin{aligned} \Phi_4 = & \frac{2}{15} + 2 \left( 1 - \frac{8}{x^2} \right) \coth^2 \frac{x}{2} \\ & + \coth^4 \frac{x}{2} - \frac{8}{3x^2} + \frac{48}{x^4} + \frac{16}{T^4} (\lambda_E + \lambda_C) \end{aligned}$$

Coefficients  $M_i$ 

$$M_1 = 2 \left( \coth \frac{x}{2} - \frac{2}{x} \right)$$

$$M_2 = \left( \coth \frac{x}{2} - \frac{6}{x} \right) \left( \coth \frac{x}{2} - \frac{2}{x} \right)$$

$$M_3 = \frac{2}{x} \left[ 1 - 3 \left( \coth \frac{x}{2} - \frac{2}{x} \right) \coth \frac{x}{2} \right]$$

Coefficients  $G_i$ 

$$G_0 = \alpha_2 + \alpha_1 \varphi_1 + \varphi_2 + 1 - (L_2 + L_1 M_1 + M_2)$$

$$\begin{aligned} G_1 = & \alpha_3 + \alpha_2 \varphi_1 + \alpha_1 \varphi_2 + \varphi_3 + (\alpha_1 + \varphi_1) \\ & - (L_2 M_1 + L_1 M_2 + M_3) \end{aligned}$$

$$\begin{aligned} G_2 = & \alpha_3 \varphi_1 + \alpha_2 \varphi_2 + \alpha_1 \varphi_3 + \varphi_4 + (L_2 + L_1 M_1 + M_2) \\ & - (L_2 M_2 + L_1 M_3) \end{aligned}$$

$$G_3 = \alpha_3 \varphi_2 + \alpha_2 \varphi_3 + \alpha_1 \varphi_4 + (L_2 M_1 + L_1 M_2 + M_3) - L_2 M_3$$

$$G_4 = \alpha_3 \varphi_3 + \alpha_2 \varphi_4 + (L_2 M_2 + L_1 M_3)$$

$$G_5 = \alpha_3 \varphi_4 + L_2 M_3$$

Coefficients  $r_i$ 

$$r_2 = (R_2)^{\frac{1}{2}}$$

$$\text{avec } \begin{cases} R_2 = \coth \frac{u}{2} \left[ \left( \frac{1}{3} - \frac{4}{u^2} \right) \coth \frac{u}{2} + \frac{8}{u^3} \right] \\ R_1 = -\frac{1}{3} + \frac{4}{u^2} - \coth^2 \frac{u}{2} \end{cases}$$

$$r_1 = (2r_2 - R_1)^{\frac{1}{2}}$$

## Références

- <sup>1</sup> RAGAZZINI, J. G. and FRANKLIN, C. *Sampled-Data Control Systems*. 1958. New York; McGraw-Hill
- <sup>2</sup> CHANG, S. S. L. Statistical design theory for digital-controlled continuous systems. *Trans. Amer. Inst. elect. Engrs II* (1958), 77
- <sup>3</sup> TOU, J. T. *Digital and Sampled Data Control Systems*. 1959. New York; McGraw-Hill
- <sup>4</sup> BERGEN, A. R. On the statistical design of linear random sampling systems. *Automatic and Remote Control*. p. 430. 1960. London; Butterworths
- <sup>5</sup> THELLIER, P. L. Optimisation et auto-optimisation des Systèmes de commande à données échantillonnées en présence de saturations. *Mem. Artillerie Franc.* T 35 (1961), 3e fasc.
- <sup>6</sup> SKLANSKY, J. On closed-form expansion for mean squares in discrete-continuous systems. *Inst. Radio Engrs., Trans. Automat. Contr.* PGAC 4 (1958), 24
- <sup>7</sup> NEWTON, G. C., GOULD, L. A. and KAISER, J. F. *Analytical Design of Linear Feedback Controls*. 1957. New York; Wiley
- <sup>8</sup> JAMES, H. M., NICHOLS, N. B. and PHILLIPS, R. S. *Theory of Servo Mechanisms*. 1947
- <sup>9</sup> LEFÈVRE, P. L'optimisation Statistique du Guidage par alignement d'un engin autopropulsé en présence de bruit. *Congrès A.G.A.R.D. sur la Stabilité et le Contrôle*. Bruxelles Avril 1961

## DISCUSSION

R. E. KING, *Electrical Engineering Dept., Queen's University, Belfast, N. Ireland*

Closed-form expressions for determining the mean squared values of signals in discrete-continuous systems have been obtained by Sklansky<sup>1</sup> and Mori<sup>2</sup>. With reference to Figure A where  $A_{ux}^*(e^{sT})$  is the 'discrete

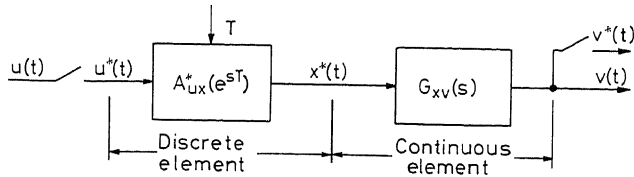


Figure A

transfer transform' in the system and  $G_{xv}(s)$  is the continuous part, the value of the mean squared output, for example, can be expressed in either of two forms:

$$\bar{v}^2 = \frac{1}{2\pi j} \oint_c \Psi_u^*(z) |A_{ux}^*(z)|^2 Z [ |G_{xv}(s)|^2 ] \frac{dz}{Tz} \quad (1)$$

due to Sklansky, and

$$\bar{v}^2 = \frac{1}{2\pi j} \oint_c \Psi_u^*(z) |A_{ux}^*(z)|^2 \left[ \int_0^1 |G_{xv}^*(z, m)|^2 dm \frac{dz}{z} \right] \quad (2)$$

due to Mori, using modified z transforms, where

$$\Psi_u^*(z) = \Phi_u^*(z) \Phi_u^*(z^{-1}) = |\Phi_u^*(z)|^2$$

is the input pulse spectral density.

In general, for physically realizable stable systems the z transform

$$Z [ |G_{xv}(s)|^2 ]$$

can be expressed in a symmetrical modulus squared form  $|\Gamma^*(z)|^2$ . Thus Sklansky's integral becomes

$$\bar{v}^2 = \frac{1}{2\pi j} \oint_c |\Phi_u^*(z) A_{ux}^*(z) \Gamma^*(z)|^2 \frac{dz}{Tz} \quad (3)$$

a rational fraction of complementary conjugate polynomials in z, i.e.

$$J_n(z) = \frac{1}{2\pi j} \oint_c \left| \frac{p_{n-1}z^{n-1} + p_{n-2}z^{n-2} + \dots + p_0}{q_n z^n + q_{n-1}z^{n-1} + \dots + q_0} \right|^2 \frac{dz}{z} \quad (4)$$

By using the bilinear transformation

$$z = \frac{1+\lambda}{1-\lambda}$$

these can be expressed in the standard form given elsewhere<sup>3</sup>.

Tables of symmetrical modulus squared z transforms  $|\Gamma^*(z)|^2$  for the most commonly found types of transforms  $G_{xv}(s)$  have been derived<sup>4</sup> for use in the analysis and optimization of amplitude dependent discrete-continuous systems. An example is

$$G_{xv}(s) = \frac{1}{s^2(s+a)}$$

for which

$$|\Gamma^*(z)|^2 = Z [ |G_{xv}(s)|^2 ] = \frac{e^{-aT}}{96 a^5} \left| \frac{(c_0 + c_1 + c_2) z^2 + 2(c_0 - c_2) z + (c_0 - c_1 + c_2)}{(z-1)^2 (z-e^{-aT})} \right|^2$$

where

$$c_0^2 = 2 \varrho_0 + 2 \varrho_1 + \varrho_2$$

$$c_2^2 = 2 \varrho_0 - 2 \varrho_1 + \varrho_2$$

$$c_1^2 = -12 \varrho_0 + 2 \varrho_2 + 2 c_0 c_2$$

$$\varrho_0 = 6 \sinh aT - 6 aT - (aT)^3$$

$$\varrho_1 = 12 aT (1 + \cosh aT)$$

$$-2 (aT)^3 (2 - \cosh aT)$$

$$-24 \sinh aT$$

$$\varrho_2 = 36 \sinh aT$$

$$+2 (aT)^3 (4 \cosh aT - 1)$$

$$-12 aT (1 + 2 \cosh aT)$$

An example of the application of the modulus squared z transform to optimum system synthesis of a fixed configuration problem is shown in Figure B. Here<sup>4</sup> contours of equal mean squared error ratio  $(\bar{e}^2/\bar{u}^2)$  of a particular sampled data system are shown for a range of values of the two free system parameters  $\omega_0$  and  $\zeta$ . The optimum point  $(\omega_0^*, \zeta^*)$  gives the minimum possible mean squared error of the particular system configuration when the sampling period and input spectrum are fixed.

However, in practical systems where non-linearity is inherent there exist certain fundamental system constraints, as for example the torque capacity of the system. Thus the movement of the parameters in the  $\omega_0 - \zeta$  plane must be constrained if a realistic optimum is to be achieved.

In order, therefore, to constrain the mean square value of a system variable, in an attempt to avoid distortion, further sets of loci of the mean squared value of the signals to be constrained may be superimposed on the mean squared error ratio loci of Figure B. Where the

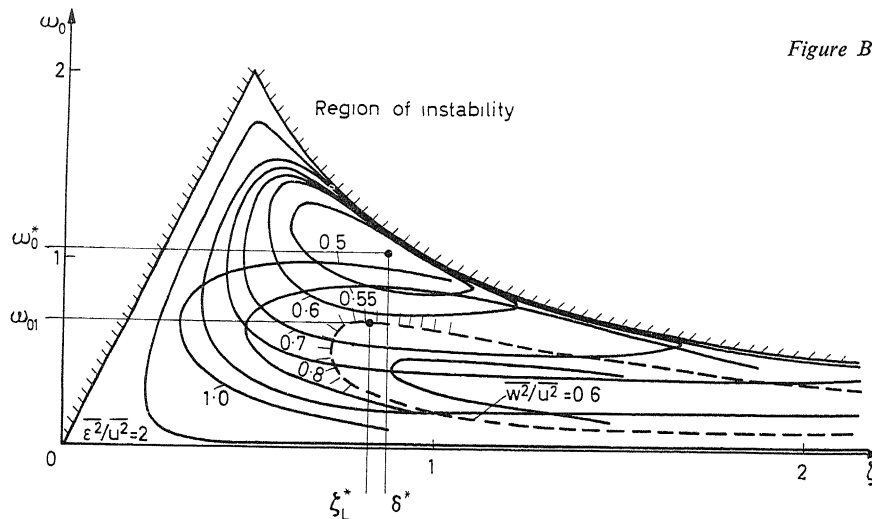


Figure B

system is subject to one constraint,  $\bar{\omega}^2 \leq \bar{\omega}_m^2$  then a 'restricted' optimum must be found on the  $\bar{\omega}_m^2$  locus. In Figure B the true optimum is as shown at  $(\omega_0^*, \zeta^*)$  and the 'restricted' optimum at  $(\omega_{0R}^*, \zeta_R^*)$ . This in effect is a graphical Lagrangian technique<sup>5</sup> and is of particular use where analytical methods are exceedingly complex.

#### References

- SKLANSKY, J. On closed-form expressions for mean squares in discrete-continuous systems. *Trans. I.R.E.*, Vol. A-4 (March 1958)
- MORI, M. Statistical treatment of sampled data control systems for actual random inputs. *Trans. A.S.M.E.*, Vol. 80 (February 1958)
- NEWTON, G. C., GOULD, L. A. and KAISER, J. F. *Analytical design of linear feedback controls*. 1957. New York; Wiley
- KING, R. E. and BROWN, W. A. Stochastic analysis of linear time-quantized control systems (Submitted for publication)
- KING, R. E. and BROWN, W. Y. Restricted optimization of a class of time quantized control systems (In preparation)

P. LEFÈVRE, *in reply*

The use of the variable  $v$  (or  $\lambda$ ) as suggested by Dr. King is exactly what we did a few years ago and now deem superfluous for the following two reasons: (1) For very complex sampled-data systems, the change of variable comes too late to be really efficient from the viewpoint of computing accuracy. With respect to the  $v$  domain optimization, it involves an appreciable number of additional operations for passing from eqn (3) to eqn (4), as also for introducing the variable  $v$ .

Furthermore, it does not make it possible to avail oneself of the simplicity of the  $v$  transforms with respect to  $z$  transforms for complex systems, as shown for example by the respective transforms of  $1/s^9$  below:

$$Z\left(\frac{1}{s^9}\right) = \frac{T^8}{8!} \cdot \frac{z}{(z-1)^9} \cdot [z^7 + 1 + 247z(1+z^5) + 4 \cdot 293z^2(1+z^3) + 15 \cdot 619z^3(1+z)]$$

$$V\left(\frac{1}{s^9}\right) = \frac{T^8}{2^9} \cdot \frac{(1-V^2)}{V^9} \cdot \left[1 - 2V^2 + \frac{6}{5}V^4 - \frac{62}{315}V^6\right]$$

(2) For reasonably simple systems the change of variable proposed is useless, because of the recent publications of Professor Jury's table of integrals, which permit a direct determination of integrals of type (4).

A detailed exposition of the  $v$  domain method of optimization will be found in Vol. 3 (in preparation) of the series 'Progress in Automatic Control Engineering' (Heywood, London), edited by Professors Macmillan, Naslin and Higgins.

B. M. BROWN, *Royal Naval College, Greenwich, London, S.E. 10*

Dr. Lefèvre has produced an interesting solution of the problem of optimization of a sampled-data linear system. He claims two advantages over the more direct method making use of the  $z$  transform and I would like to discuss these in turn.

The first advantage claimed is that by making the bilinear transformation  $v = (z-1)/(z+1)$ , the method used in the analogous problem for a continuous system becomes applicable. This is of course true, but the corresponding method for a discrete system is very similar analytically and no more difficult to apply. Indeed if it is desired to have a common method to cover the two types of system a case could be made out for taking the discrete system as basic, for in this case one is dealing with Fourier series rather than Fourier integrals, and integrals round a finite closed contour rather than along an infinite line. Results and methods for discrete systems could then be applied to continuous systems by the simple device of sampling. Provided that the various input functions of continuous time are stationary statistically the mean squared error found at the sampling instants will be equal to that at any other instant.

The bilinear transformation is frequently quoted as a device for reducing discussion of the stability of a discrete system to that of a continuous system. In fact it turns out to be rather disappointing for this purpose. Unless the degree of the characteristic equation is low the arithmetic or algebra involved in carrying out the transformation is not insignificant. As Professor Jury has reminded us earlier in this Congress, there are available direct methods for discrete systems analogous to and no more difficult to apply than those of Routh and Hurwitz for continuous systems. The bilinear transformation is not even particularly helpful in deriving one set of conditions from the other.

In general, therefore, it would appear that, other considerations being equal, it is better to use the particular method appropriate to the particular type of system. However, Lefèvre reports that he has had arithmetical difficulties when the various integrals are expressed in terms of  $z$  due to concentration of the poles of the integrands inside the unit circle. These could well justify the transformation to  $v$ . I wonder, however, whether the difficulties could be avoided by evaluating the integrals in terms of  $z$  by other methods, such as the application of Cauchy's theorem to the poles outside the unit circle or the use of Parseval's theorem. Alternatively, a table can easily be produced which gives integrals of this type directly in terms of the coefficients



of the numerator and of the partially factorized denominator of the integrand. Such a table would be analogous to that which has been calculated for continuous systems.

P. LEFÈVRE, *in reply*

The two following points in Professor Brown's comments seem to apply directly to statistical optimization.

(1) It is suggested that the statistical investigation of continuous systems could be carried out by the method developed for sampled-data systems, by providing the system with a fictitious sampler. This proposal seems to be carrying a bit too far the desire for a standardization of the method! It does not seem to agree, either, with the arguments expounded in the oral presentation of the paper.

(2) The determination of the R.M.S. values by the method of residues or by means of a table of integrals is quite interesting, as mentioned before, for simple systems, but proves impractical for very complex structures.

YA. Z. TSYPKIN, *Institute of Automation and Telemechanics, Kalachevskaya 15-A, Moscow I-53, U.S.S.R.*

In the report, the author proposes to use the well-known transformation  $Z = \frac{1-v}{1+v}$  for the statistical optimization of impulse systems, which was used for the same purpose in other papers<sup>1, 2</sup>. The application of this transformation changes the problem of statistical optimization of impulse systems to the statistical optimization of analogue systems. But using digital computers, we should again be sampling, which means that auxiliary errors will be introduced. Thus, the method of solution proposed uses unnecessary transformation of the initial equations of the impulse systems with sampling following; therefore it is easier to use the  $z$  transform (or  $D$  transformation). As for the

evaluation of loop integrals which the author considers problematical, I hope that all difficulties can be avoided by using the method of polynomial equations developed by Volgin<sup>3</sup>.

Examples of optimization of impulse systems which can easily be handled on digital computers are given in Volgin's book<sup>3</sup> and in my own book<sup>4</sup>. The latter way seems to be preferable, both with respect to the number of operations and the accuracy that can be obtained.

#### References

- <sup>1</sup> JOHNSON, G. W., LINDORF, D. P. and NORDLING, C. G. A. Extension of continuous data system design techniques to sampled-data control systems. *Trans AIEE*, Vol. 74 Pt. 2 (1955) 252-263
- <sup>2</sup> KUZIN, L. T. Proektirovanie i raschet diskretnykh avtomaticheskikh sistem. *Mashgiz* (1963)
- <sup>3</sup> VOLGIN, L. N. Elementui teorii upravlyushchikh mashin. *Mashgiz* (1962)
- <sup>4</sup> TSYPKIN, YA. Z. Teoriya lineinykh impurnykh sistem. *Mashgiz* (1963)

P. LEFÈVRE, *in reply*

I wish to thank Professor Tsyppkin for his interesting comments. I find it rather difficult to formulate my answer because I was not up to now aware of Volgin's method which does not seem to have been available at the time I carried out my work. My only information concerning the method is that it requires the transformation of the various functions involved in the form of polynomials. In this respect, it would seem to be related to the methods mentioned by Professor Brown and Dr. King. Finally I should like to ask Professor Tsyppkin whether Volgin's method has actually been applied to systems of the same order of complexity as those we mentioned.

# The Control of Two Output Dependent Processes

C. N. KERR and G. D. S. MACLELLAN

## Summary

Certain processes which require to be controlled by means of feedback loops have small perturbation static or dynamic characteristics whose parameters are functions of the unperturbed output level. The form of the response of such systems to large disturbances may be a significant design requirement. The paper describes the first results of a study of the response of controlled systems of this type, and their relation to characteristics which can be deduced from an extension of the small-perturbation roots-locus method of analysis for such processes.

Results are described for the proportional control of processes whose small-perturbation transfer functions consist of an output-dependent gain and either a simple fixed pole or a complex fixed pole-pair.

For the first controlled process the step response for large disturbances is similar to that for a first-order linear system. The time-constants of nearly equivalent linear systems are therefore derived in terms of the pole-positions of the small-perturbation roots-loci. For the second controlled process the step response for large disturbances may be stable or unstable. Criteria for stability and the form of stable response are investigated by means of *Liapunov's Direct Method*, a *phase-plane analysis*, an extension of the *roots-locus concept* to that of a *small-perturbation 'roots-surface'*, and an *approximate method of analysis*. The 'roots-surface' concept appears to be a promising method of analysis of the response of controlled processes of this class, and the investigation is being extended.

## Sommaire

Certains processus exigent une commande par boucles à réaction ont des caractéristiques statique et dynamique sous faible perturbation dont les paramètres sont des fonctions du niveau de sortie non perturbé. La forme de la réponse de tels systèmes à de fortes perturbations peut constituer une condition importante de la conception. Le rapport décrit les premiers résultats d'une étude sur la réponse de systèmes commandés de ce type, et leur relation avec des caractéristiques pouvant se déduire d'une méthode d'analyse du lieu des racines sous faible perturbation.

On décrit des résultats concernant le réglage proportionnel de 2 processus dont les fonctions de transfert, sous faible perturbation, consistent en un gain dépendant du niveau de sortie et un pôle fixe simple, ou un couple de pôles fixes complexes.

Pour le premier processus réglé, la réponse à l'échelon unité pour de fortes perturbations est semblable à celle de systèmes linéaires du premier ordre. Les constantes de temps de systèmes linéaires à peu près équivalents sont donc déduites en fonction de la position des pôles du lieu des racines sous perturbation faible.

En ce qui concerne le second processus réglé, la réponse à l'échelon unité pour une forte perturbation peut être stable ou instable. On examine alors les critères de stabilité et la forme des réponses stables au moyen de la méthode directe de Liapunov, une analyse du plan de phase, une extension du concept du lieu des racines à celui de «surface des racines» sous perturbation faible, et une méthode d'analyse approchée. Le concept de «surface des racines» paraît être une méthode prometteuse d'analyse de réponse de processus réglés de ce type, et son étude se poursuit.

## Zusammenfassung

Gewisse geregelte Systeme besitzen bei kleinen Störungen ein statisches und dynamisches Verhalten, das sich aus deren Differential-

gleichungen mit Koeffizienten, die Funktionen des ungestörten Ausgangswertes sind, ergibt. Das Übergangsverhalten solcher Systeme bei großen Störungen kann für den Entwurf von Bedeutung sein. Die Arbeit beschreibt die ersten Ergebnisse einer Untersuchung des Übergangsverhaltens solcher Regelsysteme und gibt deren Beziehungen zu Eigenschaften an, die sich als eine Erweiterung der Wurzelortskurven-Methode bei kleiner Störung betrachten läßt.

Die Ergebnisse der proportionalen Regelung von Strecken, die eine Übertragungsfunktion besitzen, die bei kleinen Störungen eine ausgangabhängige Verstärkung und entweder einen festen einfachen Pol oder ein festes komplexes Polpaar haben, sind dargestellt. Hat die Regelstrecke einen einfachen Pol, so ist bei großen Störungen die Sprungantwort ähnlich der eines linearen Systems erster Ordnung. Die Zeitkonstanten des fast äquivalenten linearen Systems werden daher als Polstellen der Wurzelortskurve bei kleinen Störungen abgeleitet. Für eine Regelstrecke mit komplexen Polpaaren kann die Sprungantwort bei großen Störungen sowohl stabil als auch instabil sein. Die Stabilitätsbedingungen und die Form des stabilen Verlaufes werden nach verschiedenen Möglichkeiten untersucht, nämlich erstens mit Hilfe der direkten Methode von Liapunov, zweitens durch die Darstellung in der Phasenebene, drittens nach einer Erweiterung der Wurzelortsmethode bei kleinen Störungen auf eine „Wurzelfläche“ und schließlich mittels einer Näherungsmethode. Der Begriff der „Wurzelfläche“ scheint eine erfolgversprechende Untersuchungsmethode für derartige Regelsysteme zu sein; die Untersuchungen werden fortgeführt.

## Introduction

A particular class of physical process is that whose input  $\theta$  and output  $\theta_0$  are related by a differential equation of the form

$$\sum_{m=1}^n (a_m + a_{m0}\theta) \theta_0^{(m)} + \left( a_0 + a_{00}\theta + \sum_{j=1}^k a_{0j}\theta^{(j)} \right) \theta_0 = b + b_0\theta + \sum_{j=1}^k b_j\theta^{(j)} \quad (1)$$

in which superscripts denote derivatives with respect to time, and any  $a$  or  $b$  may be negative or zero. As an example, the equation

$$\frac{d^2\theta_0}{dt^2} + (a_1 - \theta) \frac{d\theta_0}{dt} - \left( a_{00}\theta + \frac{d\theta}{dt} \right) \theta_0 = 0 \quad (2)$$

can represent the response of a nuclear reactor on a one-point, one delayed-neutron-group basis, where  $\theta_0$  is the neutron population and  $\theta$  the reactivity. The differential equation of such a process is characterized by the presence of products both of  $\theta_0$  with derivatives of  $\theta$ , and of derivatives of  $\theta_0$  with  $\theta$ .

The relationship between an input equilibrium level  $\theta_e$  and the corresponding output equilibrium level  $\theta_{0e}$  is obtained from eqn (1) with all derivatives equated to zero:

$$(a_0 + a_{00}\theta_e) \theta_{0e} = b + b_0\theta_e \quad (3)$$

If  $\theta$  has a small variation  $\delta\theta$  about  $\theta_e$ , and  $\theta_0$  has a resultant small variation  $\delta\theta_0$  about  $\theta_{0e}$ , linearization along with eqn (3) yields the following transfer function between  $\delta\theta$  and  $\delta\theta_0$ :

$$\frac{\delta\theta_0(p)}{\delta\theta(p)} = \frac{\sum_{j=0}^k (b_j - a_{0j}\theta_{0e}) p^j}{\sum_{m=0}^n (a_m + a_{m0}\theta_e) p^m} \quad (4)$$

the form of which is a function of  $\theta_e$  and  $\theta_{0e}$ . Use of eqn (3) to eliminate  $\theta_e$  produces a small-signal transfer function which is a function of the unperturbed output level  $\theta_{0e}$  only. Any process whose dynamic response can be described by eqn (1) can thus, in a restricted sense, be described as an 'output-dependent process'.

For a feedback control loop round a process of this kind, such that  $\theta$  is given by  $(\theta_i - \theta_0)$  where  $\theta_i$  is the input signal to the system, the shape of the roots-loci for small-signal response is a function of  $\theta_{0e}$ . A 'roots-surface' can therefore be constructed in a three-dimensional space whose axes represent the real and the imaginary frequency components and  $\theta_{0e}$ . From the shape of this surface, the small-signal response of the closed-loop system can be directly deduced for all values of  $\theta_{0e}$  and of loop gain.

In this paper two special cases are investigated for the relations which exist between the characteristics of the roots-surface and the response of the closed-loop system for large variations in its state. On the basis of these relations design procedures for the control of processes of this class are being developed which are similar to those used in the roots-locus method of synthesis for linear systems and for systems which contain single non-linear elements.

### The Control of a Process Whose Transfer Function Consists of a Variable Gain and a Simple Fixed Pole

Such a process is characterized by the relation

$$\frac{d\theta_0}{dt} + (a_0 + a_{00}\theta) \theta_0 = b_0\theta$$

which may be written as

$$T \frac{d\theta_0}{dt} + (1 - \lambda K\theta) \theta_0 = K\theta$$

where  $b_0/a_0 = K$ ,  $a_{00}/b_0 = -\lambda$ , and  $1/a_0 = T$ , and  $K$ ,  $T$ , and  $\lambda$  are all positive. Figure 1 is a form of block diagram for the closed-loop system around the process, which shows that it can be considered to consist of a first-order lag preceded by a gain element whose instantaneous value depends on the instantaneous

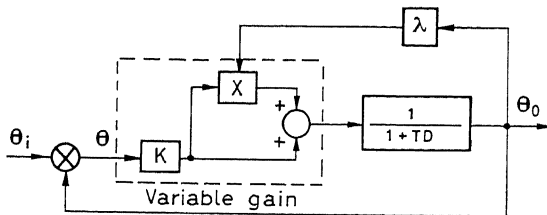


Figure 1. Block diagram for process with variable gain and simple fixed pole

value of the output from the lag. In the subsequent consideration of this system a restriction  $\theta_0 > -(1/\lambda)$  is imposed, so that the process gain has only positive values.

For any small perturbation  $\delta\theta_i$  from the equilibrium state  $(\theta_{ie}, \theta_{0e})$ , the transmission from  $\delta\theta_i$  to  $\delta\theta_0$  obtained from eqns (3) and (4) is shown in Figure 2.

### Roots Surface

To construct the roots-surface in the  $(\sigma, j\omega, \lambda\theta_{0e})$  space (Figure 3), the path of the open-loop pole  $-1/T(1 + \lambda\theta_{0e})$  is drawn first. This lies entirely in the  $(\sigma, \lambda\theta_{0e})$  plane, so that the roots-surface in this case is simply the portion of the  $(\sigma, \lambda\theta_{0e})$  plane on the negative side of the open-loop pole path. The root

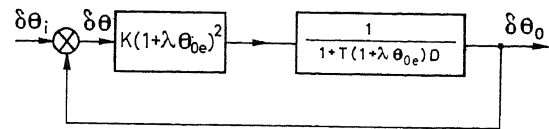


Figure 2. Block diagram for small variations in the process of Figure 1

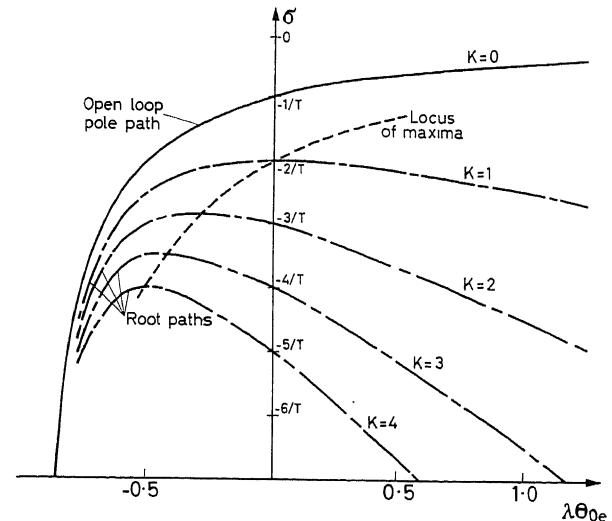


Figure 3. Roots-surface and root paths for first-order lag process

paths lie in the surface at a distance  $K/T(1 + \lambda\theta_{0e})$  from the open-loop pole path; a few are drawn in Figure 3 for different values of  $K$ . The equation of a root path is

$$\sigma = -\frac{1}{T} \frac{1 + K(1 + \lambda\theta_{0e})^2}{1 + \lambda\theta_{0e}} \quad (5)$$

where  $\sigma$  is the position of the root; each path has a maximum at  $\lambda\theta_{0e} = (1/K)^{1/2} - 1$ , for which

$$\sigma = -\frac{2}{T(1 + \lambda\theta_{0e})}$$

and the locus of maxima is indicated on Figure 3.

### Large Step Response

Consider a large step change in  $\theta_i$  applied to the system at  $t = 0$ . The initial equilibrium values of  $\theta_i$  and  $\theta_0$  for  $t < 0$ , denoted by  $\theta_{i0}$  and  $\theta_{00}$  respectively, are related by

$$K\theta_{i0}(1 + \lambda\theta_{00}) = \theta_{00}(1 + K(1 + \lambda\theta_{00}))$$

and, for  $t \geq 0$ ,  $\theta_i = \theta_{if}$ . The final equilibrium value for  $\theta_0$ , denoted by  $\theta_{0f}$ , is given by

$$K\theta_{if}(1 + \lambda\theta_{0f}) = \theta_{0f}(1 + K(1 + \lambda\theta_{0f}))$$

The system responds according to the equation

$$T \frac{d\theta_0}{dt} + (1 + K - \lambda K\theta_{if})\theta_0 + \lambda K\theta_0^2 = K\theta_{if}$$

This may be solved by separating the variables to give

$$\theta_0 = \theta_{0f} \frac{A - B e^{-Ct}}{A + E e^{-Ct}} \quad (6)$$

in which

$$A = 1 + K(1 + \lambda\theta_{0f})(1 + \lambda\theta_{00})$$

$$B = \left(1 - \frac{\theta_{00}}{\theta_{0f}}\right)(1 + K(1 + \lambda\theta_{0f}))$$

$$C = \frac{1 + K(1 + \lambda\theta_{0f})^2}{T(1 + \lambda\theta_{0f})}$$

$$E = \lambda K(\theta_{0f} - \theta_{00})(1 + \lambda\theta_{0f})$$

To correlate this response with the characteristics of the roots-surface, eqn (6) may be approximated by a single exponential relation

$$\theta_0 = \theta_{0f} - (\theta_{0f} - \theta_{00}) e^{-t/T_{eq}} \quad (7)$$

where  $T_{eq}$  is chosen to give in some sense an optimum representation. If the chosen criterion is that the integral  $\int_0^\infty (\theta_{0f} - \theta_0) dt$  should be the same for the exact solution (6) as for the approximate solution (7) (cf. Nechleba<sup>1</sup>), the following result is obtained:

$$\sigma_{eq} = -\frac{1}{T_{eq}} = -\frac{\lambda K(\theta_{00} - \theta_{0f})}{T \log \left( 1 + \frac{\lambda K(1 + \lambda\theta_{0f})(\theta_{00} - \theta_{0f})}{1 + K(1 + \lambda\theta_{0f})^2} \right)} \quad (8)$$

With reference to eqn (5), if the small-perturbation pole at  $\theta_{0e} = \theta_{0f}$  is denoted by  $\sigma_f$ , eqn (8) becomes

$$\sigma_{eq} = -\frac{\lambda K(\theta_{00} - \theta_{0f})}{T \log \left( 1 + \frac{\lambda K(\theta_{0f} - \theta_{00})}{T\sigma_f} \right)} \quad (9)$$

This expresses the equivalent single pole as a function of the initial and final equilibrium values of  $\theta_0$  and of the small-perturbation pole position at  $\theta_{0f}$ .

Furthermore, if  $|T\sigma_f| > |\lambda K(\theta_{0f} - \theta_{00})|$ , so that the expansion  $\log(1 + x) = x - \frac{1}{2}x^2 + \frac{1}{3}x^3 - \dots$  converges, then

$$\sigma_{eq} = \sigma_f \left/ \left( 1 - \frac{1}{2}u + \frac{1}{3}u^2 - \dots \right) \right.$$

where  $u = \lambda K(\theta_{0f} - \theta_{00})/T\sigma_f$ , which may be a preferable form. Morgan<sup>2</sup> has suggested an alternative criterion for  $T_{eq}$  which, unlike the first, does not require the exact solution for  $\theta_0$ . It is

$$T_{eq} = \frac{\theta_{0f} - \theta_{00}}{2 \left( \frac{d\theta_0}{dt} \right)_{av}}, \quad \text{where} \quad \left( \frac{d\theta_0}{dt} \right)_{av} = \frac{1}{\theta_{0f} - \theta_{00}} \int_{\theta_{00}}^{\theta_{0f}} \frac{d\theta_0}{dt} \cdot d\theta_0$$

This gives the result

$$\sigma_{eq} = \frac{1}{3}\sigma_f + \frac{2}{3}\sigma_0 + \frac{2}{3} \frac{\lambda(\theta_{0f} - \theta_{00})}{T(1 + \lambda\theta_{0f})(1 + \lambda\theta_{00})} \quad (10)$$

where  $\sigma_0$  denotes the small-perturbation pole at  $\theta_{0e} = \theta_{00}$ .

The accuracy of representation by the equivalent time constants is illustrated in Figure 4 by the responses for  $K = 4$ , with  $\lambda\theta_{00} = \pm 0.5$  and  $\lambda\theta_{0f} = \mp 0.5$  respectively; in each case the movement of the small-perturbation pole is considerable. It appears that the use of either equivalent pole is of adequate accuracy for engineering purposes in estimating the nature of large transients from the roots surface.

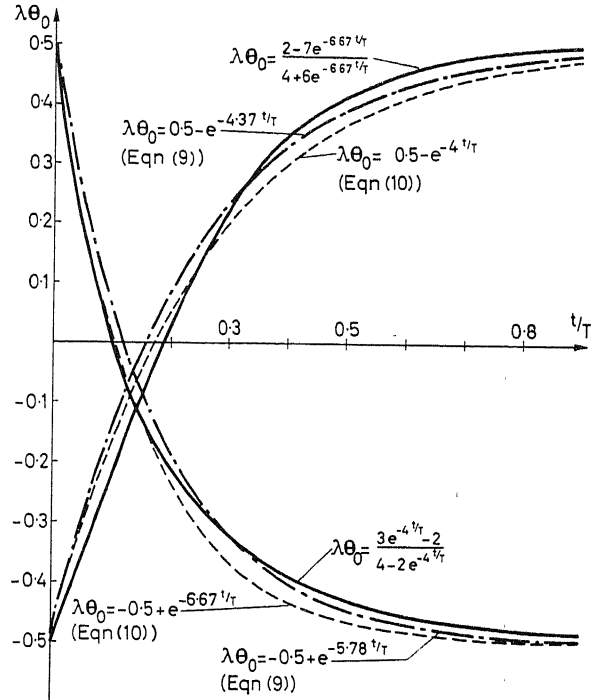


Figure 4. Typical responses of the closed loop round the first-order lag process, with corresponding approximations from eqns (9) and (10)

### The Control of a Process whose Transfer Function Consists of a Variable Gain and a Complex Fixed Pole-pair

This process is characterized by the relation

$$\frac{d^2\theta_0}{dt^2} + a_1 \frac{d\theta_0}{dt} + (a_0 + a_{00}\theta_0)\theta_0 = b_0\theta$$

which may be written

$$\frac{d^2\theta_0}{dt^2} + 2\zeta\omega_n \frac{d\theta_0}{dt} + \omega_n^2(1 - \lambda K\theta_0)\theta_0 = \omega_n^2 K\theta$$

where

$$a_1 = 2\zeta\omega_n, \quad a_0 = \omega_n^2, \quad a_{00} = -\lambda K\omega_n^2, \quad \text{and} \quad b_0 = \omega_n^2 K,$$

and  $K$ ,  $\omega_n$ ,  $\zeta$ , and  $\lambda$  are all positive. Figure 5 is a form of block diagram for the closed-loop control system around the process, which shows that it can be considered to consist of a second order lag preceded by a gain element whose instantaneous value depends on the instantaneous value of the output from the lag.

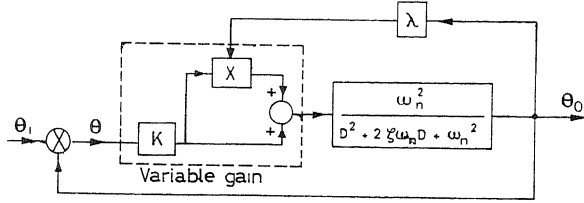


Figure 5. Block diagram for process with variable gain and complex fixed pole-pair

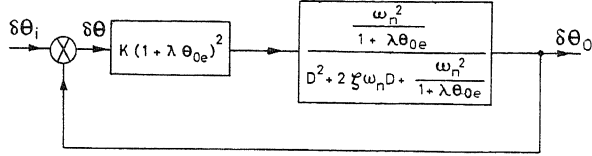


Figure 6. Block diagram for small variations in the process of Figure 5

No restriction is imposed on  $\theta_0$  for this system, so that the process gain may at times be negative.

For any small perturbation  $\delta\theta_i$  from the equilibrium state  $(\theta_{ie}, \theta_{0e})$ , the transmission from  $\delta\theta_i$  to  $\delta\theta_0$  obtained from eqns (3) and (4) is shown in Figure 6. The small-perturbation natural frequency  $\omega_e$  and damping  $\zeta_e$  is defined by the relations

$$\omega_e^2 = \omega_n^2 / (1 + \lambda\theta_{0e}) \quad \text{and} \quad \zeta_e = \frac{\omega_n \cdot \zeta}{\omega_e}$$

#### Roots Surface

To construct the roots-surface (Figure 7), use is made of non-dimensional coordinates  $\sigma/\omega_n$ ,  $j\omega/\omega_n$  and  $\lambda\theta_{0e}$ . The positions  $P (= \sigma + j\omega)$  of the two open-loop poles are given by

$$P = -\zeta_e \omega_e \pm \omega_e (\zeta_e^2 - 1)^{\frac{1}{2}} = -\zeta \omega_n \pm \omega_n \left( \zeta^2 - \frac{1}{1 + \lambda\theta_{0e}} \right)^{\frac{1}{2}}$$

For  $-1 < \lambda\theta_{0e} < (1/\zeta^2 - 1)$ , the real part  $-\zeta\omega_n$  of  $P$  is constant, and the imaginary part  $\pm j\omega_n (1/(1 + \lambda\theta_{0e}) - \zeta^2)^{\frac{1}{2}}$  tends to infinity as  $\lambda\theta_{0e}$  tends to  $-1$ ; and for  $\lambda\theta_{0e} \geq (1/\zeta^2 - 1)$  or  $\lambda\theta_{0e} < -1$ , the imaginary part of  $P$  is zero. Figure 7 is drawn for  $\zeta = 0.707$ , with the result that the roots-surface consists of:

(a) for  $-1 < \lambda\theta_{0e} < 1$ , the portions of the  $\sigma/\omega_n = -\zeta$  plane 'above and below' the open-loop pole paths;

(b) for  $\lambda\theta_{0e} > 1$ , the whole of the  $\sigma/\omega_n = -\zeta$  plane together with the portion of the  $j\omega/\omega_n = 0$  plane between the open-loop pole paths; and

(c) for  $\lambda\theta_{0e} < -1$ , the portions of the  $j\omega/\omega_n = 0$  plane 'to the left' and 'to the right' of the open-loop pole paths.

Figure 8 is complementary to Figure 7, presenting the parts of the roots-surface in the two normal planes separately. The positions of the closed-loop poles are

$$-\zeta\omega_n \pm \omega_n \left( \zeta^2 - \frac{1 + K(1 + \lambda\theta_{0e})^2}{1 + \lambda\theta_{0e}} \right)^{\frac{1}{2}} \quad (11)$$

The root-paths, a few of which are drawn in Figures 7 and 8, lie in the real plane for  $\lambda\theta_{0e} < -1$ ; if  $K \leq \zeta^4/4$  the paths lie in the real plane for

$$-1 + \frac{1}{2K} [\zeta^2 - (\zeta^4 - 4K)^{\frac{1}{2}}] < \lambda\theta_{0e} < -1 + \frac{1}{2K} [\zeta^2 + (\zeta^4 - 4K)^{\frac{1}{2}}]$$

but for  $K > \zeta^4/4$  the paths do not meet the real plane anywhere.

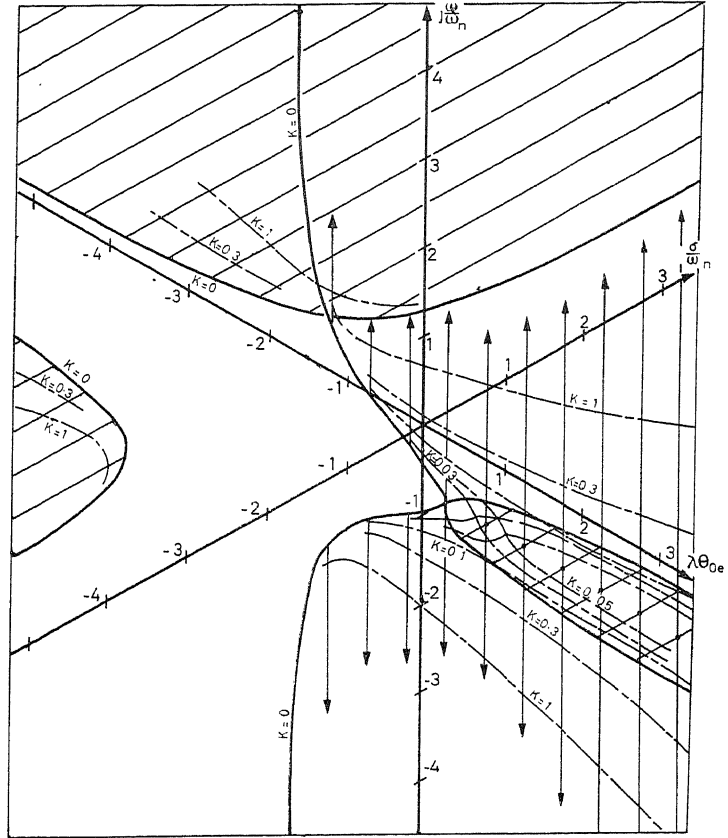


Figure 7. Roots-surface and root paths for second-order lag process, with  $\zeta = 0.707$

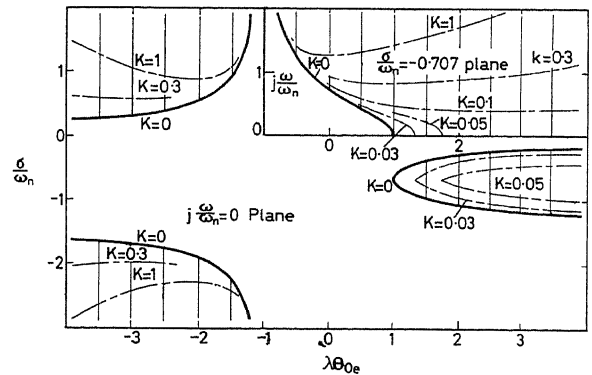


Figure 8. Two-plane presentation of roots-surface of Figure 7. (The negative half of the  $\sigma/\omega_n = -0.707$  plane is the mirror image in the  $\lambda\theta_{0e}$  axis of the positive half, and is omitted)

#### Large Step Response

Consider again a large step change in  $\theta_i$  applied to the system at  $t = 0$  for an initial equilibrium state  $(\theta_{i0}, \theta_{00})$  defined by

$$K\theta_{i0}(1 + \lambda\theta_{00}) = \theta_{00}(1 + K(1 + \lambda\theta_{00}))$$

The system responds to a disturbance  $\theta_i = \theta_{if}$  for  $t > 0$  according to the equation

$$\frac{d^2\theta_0}{dt^2} + 2\zeta\omega_n \frac{d\theta_0}{dt} + \omega_n^2(1 + K - \lambda K\theta_{if})\theta_0 + \lambda\omega_n^2 K\theta_0^2 = \omega_n^2 K\theta_{if} \quad (12)$$

This may be written more generally with  $\Theta_0 = \lambda\theta_0$  and  $\Theta_{if} = \lambda\theta_{if}$ :

$$\frac{d^2\Theta_0}{dt^2} + 2\zeta\omega_n \frac{d\Theta_0}{dt} + \omega_n^2(1+K-K\Theta_{if})\Theta_0 + \omega_n^2 K\Theta_0^2 = \omega_n^2 K\Theta_{if} \quad (13)$$

It is first necessary to investigate the stability of the system. For this Liapunov's Direct Method is suitable, with eqn (13) rewritten in the form

$$\begin{aligned} \dot{\Theta}_1 &= \Theta_2 \\ \dot{\Theta}_2 &= \omega_n^2 K\Theta_{if} - 2\zeta\omega_n\Theta_2 - \omega_n^2(1+K-K\Theta_{if})\Theta_1 - \omega_n^2 K\Theta_1^2 \end{aligned} \quad (14)$$

where  $\Theta_1 \equiv \Theta_0$  and  $\Theta_2 \equiv d\Theta_0/dt$ . The system has two critical points for which  $\dot{\Theta}_1 = 0$  and  $\dot{\Theta}_2 = 0$ , namely,<sup>†</sup>

$$(A) \quad \Theta_1 = A_1, \Theta_2 = 0, \quad \text{and} \quad (B) \quad \Theta_1 = A_2, \quad \Theta_2 = 0$$

where  $A_1$  and  $A_2$  are the roots (both real and distinct) of the quadratic

$$K\Theta_1^2 + (1+K-K\Theta_{if})\Theta_1 - K\Theta_{if} = 0$$

and  $A_1 > A_2$ .

*Point A*

The origin is transferred to *point A* by the transformation

$$\Theta'_1 = \Theta_1 - A_1, \quad \Theta'_2 = \Theta_2 \quad (15)$$

so that eqn (14) becomes

$$\begin{aligned} \dot{\Theta}'_1 &= \Theta'_2 \\ \dot{\Theta}'_2 &= -\omega_n^2(1+K-K\Theta_{if}+2KA_1)\Theta'_1 - 2\zeta\omega_n\Theta'_2 - \omega_n^2 K(\Theta'_1)^2 \end{aligned}$$

with the new origin now a critical point. The characteristic roots  $\alpha_1$  and  $\alpha_2$  of the linear approximation satisfy

$$\begin{vmatrix} -\alpha & 1 \\ -\omega_n^2(1+K-K\Theta_{if}+2KA_1) & -2\zeta\omega_n-\alpha \end{vmatrix} = 0$$

from which  $\alpha_1 + \alpha_2 = -2\zeta\omega_n$ , which is negative, and  $\alpha_1\alpha_2 = \omega_n^2(1+K-K\Theta_{if}+2KA_1)$ , which is positive. Thus both  $\alpha_1$  and  $\alpha_2$  have negative real parts, and *point A* is asymptotically stable.

*Point B*

The origin is transferred to *point B* by the transformation

$$\Theta''_1 = \Theta_1 - A_2, \quad \Theta''_2 = \Theta_2$$

Eqn (14) becomes

$$\begin{aligned} \dot{\Theta}''_1 &= \Theta''_2 \\ \dot{\Theta}''_2 &= -\omega_n^2(1+K-K\Theta_{if}+2KA_2)\Theta''_1 - 2\zeta\omega_n\Theta''_2 - \omega_n^2 K(\Theta''_1)^2 \end{aligned} \quad (16)$$

Since  $(1+K-K\Theta_{if}+2KA_2) < 0$ , the characteristic roots of the linear approximation are real and of opposite sign, so that *point B* is a saddle point.

### Stability and the Small-perturbation Roots-surface

To each value of  $\theta_{if}$  there correspond two possible equilibrium states ( $\theta_{if}, \theta_{of}$ ) given by the relation

$$K\theta_{if}(1+\lambda\theta_{of}) = \theta_{of}(1+K(1+\lambda\theta_{of})) \quad (17)$$

as shown in *Figure 9*. From the foregoing discussion of stability it follows that the higher value of  $\theta_{of}$  is stable, and the lower is unstable.

This result could also have been deduced from the roots surface of *Figure 7*, since for  $\lambda\theta_{of} < -1$  there is a small-perturbation closed-loop pole in the 'right-half space', while for  $\lambda\theta_{of} > -1$  the small-perturbation responses are stable. However, the complete analysis given above provides the necessary justification for the use of the roots-surface in this way.

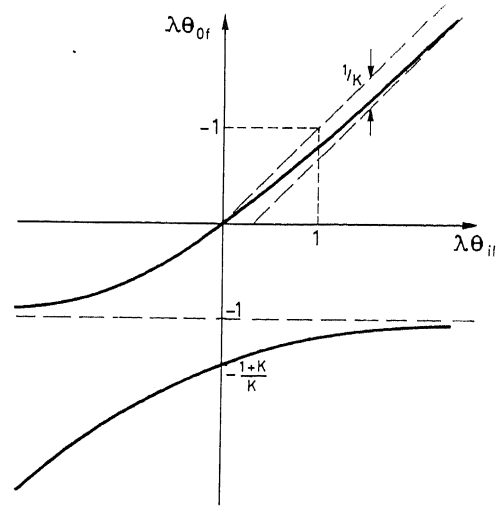


Figure 9. Dependence of  $\lambda\theta_{of}$  on  $\lambda\theta_{if}$

### Bounds of Stability

Consider now what bounds exist on  $\theta_{if}$  for a given initial equilibrium condition (either stable or unstable) to ensure that the response of  $\theta_0$  is asymptotically stable to the greater corresponding value of  $\theta_{of}$ . The problem is to find a suitable Liapunov function for system (15). This by use of (17) may be written

$$\begin{aligned} \dot{\Theta}'_1 &= \Theta'_2 \\ \dot{\Theta}'_2 &= -\omega_n^2 R\Theta'_1 - 2\zeta\omega_n\Theta'_2 - \omega_n^2 K(\Theta'_1)^2 \end{aligned}$$

where

$$R = [(K\Theta_{if} + K - 1)^2 + 4K]^{\frac{1}{2}}$$

In terms of  $\phi_1 = K\Theta'_1/R$  and  $\phi_2 = K\Theta'_2/\omega_n R^{3/2}$  this becomes

$$\begin{aligned} \dot{\phi}_1 &= \omega_n \sqrt{R}\phi_2 \\ \dot{\phi}_2 &= -\omega_n(\sqrt{R}\phi_1 + 2\zeta\phi_2 + \sqrt{R}\phi_1^2) \end{aligned} \quad (18)$$

with critical points at *A*:  $\phi_1 = \phi_2 = 0$ , and at *B*:  $\phi_1 = -1$ ,  $\phi_2 = 0$ , indicated in *Figure 10*.

Consider the function  $V_1 = 3\phi_2^2 + 3\phi_1^2 + 2\phi_1^3$ , for which  $\dot{V}_1 = -12\zeta\omega_n\phi_2^2$  (cf. Lasalle and Lefschetz<sup>3</sup>);  $\dot{V}_1$ , being zero on the  $\phi_1$  axis and negative elsewhere, is negative semi-definite.  $V_1$  is symmetric about the  $\phi_1$  axis, and, in the infinite half-plane for which  $\phi_1 > -3/2$ , it is positive except at the

origin where it is zero. The closed region  $\Omega_1$  formed by the indicated part of the  $V_1 = 1$  contour is one of asymptotic stability to the origin, because within it (a)  $V_1$  is positive (except at the origin where it is zero), (b)  $\dot{V}_1 \leq 0$ , and (c)  $\dot{V}_1$  is not identically zero along any trajectory of the system (cf. Lasalle and Lefschetz<sup>3</sup>).

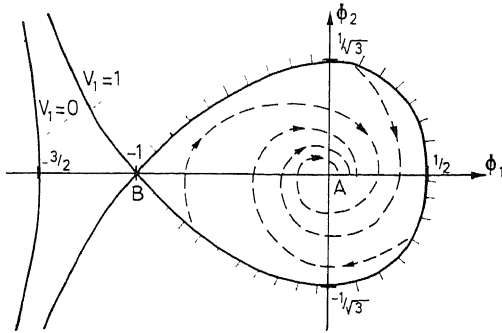


Figure 10. The region of asymptotic stability  $\Omega_1$  corresponding to  $V_1$

From the extent of  $\Omega_1$  it is evident that the restrictions on the magnitude of the input step for stable response, for particular values of  $\theta_{i0}$  and  $K$ , are  $-1 < K\theta'_{i0}/R < \frac{1}{2}$ . Corresponding to  $\theta_{00} = -1$ , the critical value of  $\theta'_{i0}$  is  $(\theta_{i0} - A_1) = (1 - K - K\theta_{if} - R)/2K$ . This critical value lies within this range, so two separate ranges must be considered:

- (a)  $(1 - K - K\theta_{if} - R)/2K < \theta'_{i0} < R/2K$  and
- (b)  $-R/K < \theta'_{i0} < (1 - K - K\theta_{if} - R)/2K$

These correspond to initial equilibrium states that are stable or unstable respectively.

(a) Using (17), the condition for a stable response from a stable initial equilibrium state is

$$K\theta_{if} + 2[(K\theta_{if} + K - 1)^2 + 4K]^{\frac{1}{2}} > K\theta_{i0} + [(K\theta_{i0} + K - 1)^2 + 4K]^{\frac{1}{2}}$$

The corresponding allowable values of  $\theta_{if}$  for stable response are shown in Figure 11.

(b) Using (17), the condition for stable response from an unstable initial equilibrium state is

$$K\theta_{if} - [(K\theta_{if} + K - 1)^2 + 4K]^{\frac{1}{2}} < K\theta_{i0} - [(K\theta_{i0} + K - 1)^2 + 4K]^{\frac{1}{2}}$$

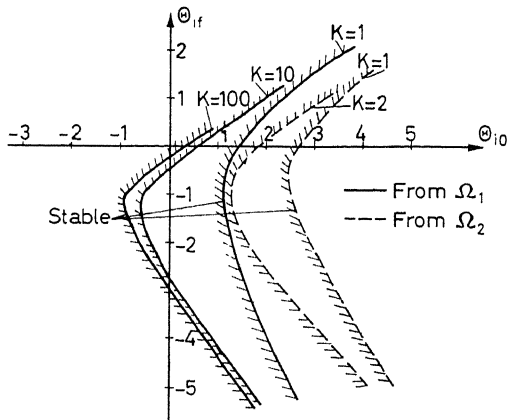


Figure 11. Bounds on  $\theta_{if}$  for stable response deduced from  $\Omega_1$  and  $\Omega_2$  (Stability exists to the left of each boundary)

This reduces to the simple inequality  $\theta_{if} < \theta_{i0}$  for all values of  $\theta_{i0}$  and positive  $K$ . Thus, in this case, the application of a negative input step results in a final stable equilibrium state.

#### Less Restrictive Bounds

The bounds on  $\theta_{if}$  derived from  $V_1$  are too restrictive, because they define ranges of allowable values for  $\theta_{if}$  which are in some cases less than are actually permissible. As a possible second Liapunov function we consider

$$V_2 = A\phi_1^2 + B\phi_2^2 + C\phi_1\phi_2 + D\phi_1^2\phi_2 + E\phi_1\phi_2^2 + F\phi_2^3 + G\phi_1^4$$

in which the values of the coefficients  $A, B, \dots$  are to be determined. To produce a form for  $\dot{V}_2$  which is symmetric about both axes, one chooses

$$A = 2(3N + 8), \quad B = 5N + 8, \quad C = \sqrt{N}(N + 8),$$

$$D = -\sqrt{N}(N + 8), \quad E = -4N, \quad F = -2/3 N^{3/2},$$

and  $G = -2N$ , where  $N = R/\zeta^2$ .

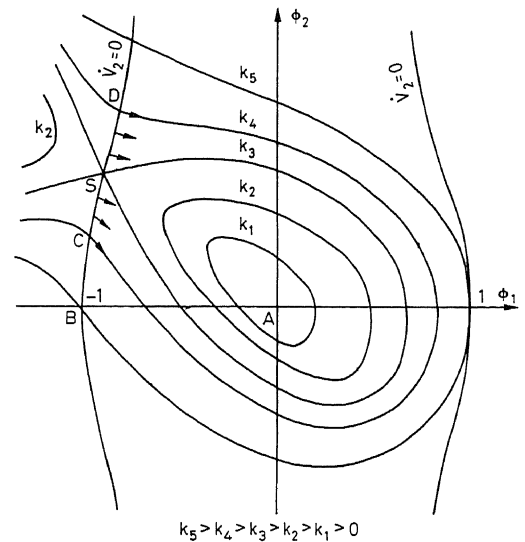


Figure 12. Sketch of contours of  $V_2$  and of  $\dot{V}_2 = 0$  for a value of  $N < 6 + 2\sqrt{17}$

$$\begin{aligned} \dot{V}_2 = \zeta \omega_n [ & (N^2 - 12N - 32)\phi_2^2 \\ & + N(N + 8)(\phi_1^2 - 1)\phi_1^2 + 2N^2\phi_1^2\phi_2^2 ] \end{aligned}$$

so that both  $V_2$  and  $\dot{V}_2$  depend on the parameter  $N$  (whereas  $V_1$  has no parametric dependence).

For values of  $N$  less than  $(6 + 2\sqrt{17}) = 14.25$ , there is an infinitely long region straddling the  $\phi_2$  axis for which  $\dot{V}_2$  is negative, except at the origin where it is zero; this lies between the two branches of the  $\dot{V}_2 = 0$  curve which intersect the  $\phi_1$  axis perpendicularly at  $\pm 1$ . The character of  $V_2$  must be examined to discover if any contour forms a closed region within the region of negative  $\dot{V}_2$ . If this exists, a region of asymptotic stability with less restrictive bounds than  $\Omega_1$  may result.

Figure 12 shows the character of  $V_2$  for a value of  $N$  less than 8.0 approximately. Up to a limiting value of  $C = k_3$ , contours  $V_2 = C$  form closed regions within the region of

negative  $\dot{V}_2$ , so that the region within  $V_2 = k_3$  is one of asymptotic stability; however, this new bound represents little, if any, improvement over the original one. But because the saddle point  $S$  of the  $V_2$  contours lies on the left-hand branch of  $\dot{V}_2 = 0$ , as indicated, it is possible, by combining Liapunov's criteria with a deduction from eqn (18), to prove the existence of a larger region of stability.

Consider the region formed by  $V_2 = k_4$  and closed by the portion  $CD$  of the curve  $\dot{V}_2 = 0$ . Throughout this region  $\dot{V}_2$  is negative, and no trajectory can leave the region across  $V_2 = k_4$ . At any point on  $CD$  where  $\dot{V}_2 = 0$ , a trajectory must run tangent to the  $V_2$  contour through the point. Because  $CD$  lies in the region  $\phi_2 > 0$ , from eqn (18)  $\dot{\phi}_1 > 0$ ; hence, every trajectory from  $CD$  has a component of velocity in the positive  $\phi_1$  direction. From inspection of the directions of the  $V_2$  contours relative to  $CD$ , it is evident that all the trajectories cross  $CD$  into the region  $\Omega_2$ , which is one of asymptotic stability.

For  $N < 8.0$  approximately, the contour  $V_2 = k_5$  through  $B$  provides the limiting contour. The condition for a stable response from an initial stable equilibrium state is therefore

$$K\Theta_{if} + 3[(K\Theta_{if} + K - 1)^2 + 4K]^{\frac{1}{2}} > K\Theta_{i0} + [(K\Theta_{i0} + K - 1)^2 + 4K]^{\frac{1}{2}}$$

The allowable values of  $\Theta_{if}$  arising from this expression, which are valid for  $N < 8.0$  approximately (i.e.  $[K(\Theta_{if} + 1) - 1]^2 < 4.0(16\zeta^4 - K)$  approximately), are also shown in Figure 11.

For  $6 + 42\sqrt{17} > N > 8.0$  approximately, the contour through  $B$  is unsuitable because it crosses the right-hand branch of the  $\dot{V}_2 = 0$  curve for a negative value of  $\phi_2$ . In this range of  $N$ , one of the intermediate contours between  $k_3$  and  $k_5$  which does not cross the branch in this way may be used to give improved bounds.

### The Phase Portrait

As eqn (13) is a second order differential equation, its solutions can be portrayed in the phase plane formed by the coordinates  $\lambda\theta_0$  and  $\lambda\dot{\theta}_0$ , to display their stability in a compact form. For a set of values of  $\zeta$ ,  $\omega_n$  and  $K$ , there is a different portrait for each value of  $\lambda\theta_{if}$ . However, if a linear transformation on  $\lambda\theta_0$  is employed to place  $A_1$  at  $(1, 0)$  and  $A_2$  at  $(-1, 0)$ , the work involved in drawing portraits for different  $\lambda\theta_{if}$  is much reduced. The transformation also gives the lower bound on  $\lambda\theta_{i0}$  for stable response, but does not give the upper bound, unlike the Liapunov function analysis.

To give a better appreciation of the nature of  $\Omega_1$  and  $\Omega_2$  in a particular case, the method of isoclines has been used to derive the phase portrait shown in Figure 13 for the process with  $\zeta = 0.707$ ,  $\omega_n = 1$ , and  $K = 1$ , and for  $\lambda\theta_{if} = 2$ . The two regions are superimposed; whilst  $\Omega_2$  represents a large improvement over  $\Omega_1$ , it still underestimates the actual upper bound that exists on  $\lambda\theta_{i0}$ .

### Approximate Estimation of Transient Response

It is known that  $V$  can give an estimate of the largest time-constant of the response, from the figure of merit

$$\eta = (-\dot{V}/V)_{min}$$

where  $1/\eta$  represents the largest time constant. However, a finite value for  $\eta$  is only achieved if  $\dot{V}$  is negative definite through-

out the region of asymptotic stability. In both  $\Omega_1$  and  $\Omega_2$ ,  $\dot{V}$  is only negative semi-definite, so no estimate of the time behaviour of the solutions is possible by this method.

An attempt to obtain an analytical solution to eqn (13) in closed form through a transformation listed by Murphy<sup>4</sup> has been unsuccessful. An approximate solution has been obtained following Grensted<sup>5</sup> by assuming that

$$\Theta_0 = \Theta_{0f} + a \sin \psi$$

where  $a(t)$  is the 'amplitude' and  $\psi(t)$  the 'phase' of  $\Theta_0$ .

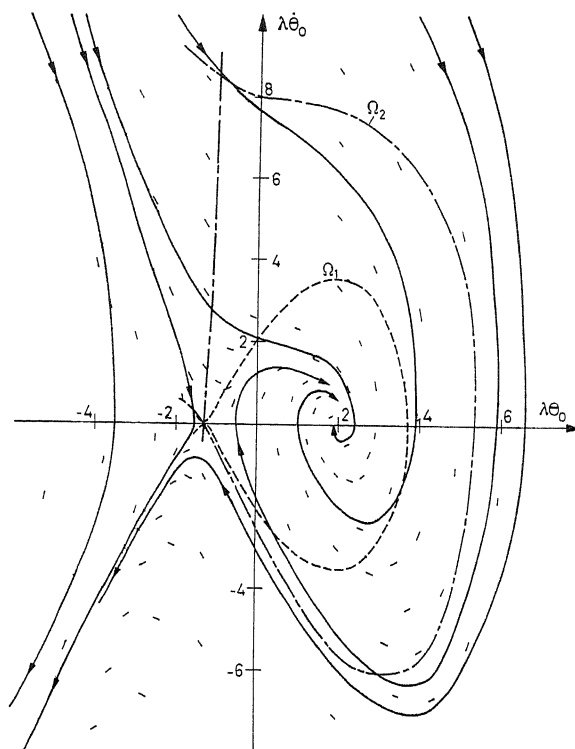


Figure 13. Phase portrait of controlled process with  $\zeta = 0.707$ ,  $\omega_n = 1$ ,  $K = 1$ , and  $\lambda\theta_{if} = 2$

We define

$$\mu = -\dot{a}/a \text{ or } a = e^{-\int_0^t \mu dt}$$

and

$$\omega = \dot{\psi} \text{ or } \psi = \int_0^t \omega dt$$

Then

$$\Theta_0 = \Theta_{0f} + A e^{-\int_0^t \mu dt} \sin(\int_0^t \omega dt + \phi)$$

where  $A$  and  $\phi$  are arbitrary constants.

Substitution of this form for  $\Theta_0$  in eqn (13) gives:

$$\begin{aligned} & \left[ \mu^2 - \omega^2 - \dot{\mu} - 2\zeta\omega_n\mu + \omega_n^2 \frac{1 + K(1 + \Theta_{0f})^2}{1 + \Theta_{0f}} \right] \sin(\int_0^t \omega dt + \phi) \\ & + [\dot{\omega} - 2\mu\omega + 2\zeta\omega_n\omega] \cos(\int_0^t \omega dt + \phi) \\ & + \frac{1}{2} \omega_n^2 K A e^{-\int_0^t \mu dt} (1 - \cos 2(\int_0^t \omega dt + \phi)) = 0 \end{aligned}$$



If the third term which decreases with time is ignored,

$$\omega^2 = \omega_n^2 \frac{1 + K(1 + \Theta_{of})^2}{1 + \Theta_{of}} + \mu^2 - \dot{\mu} - 2\zeta\omega_n\mu$$

and

$$\mu = \zeta\omega_n + \frac{1}{2} \frac{d(\omega^2)/dt}{\omega^2}$$

An iteration procedure for  $\omega^2$  and  $\mu$  follows, which converges:

$$\begin{aligned} \omega_0^2 &= \mu_0 = 0 \\ \omega_1^2 &= \omega_n^2 \frac{1 + K(1 + \Theta_{of})^2}{1 + \Theta_{of}} : \mu_1 = \zeta\omega_n \\ \omega_2^2 &= \omega_n^2 \left( \frac{1 + K(1 + \Theta_{of})^2}{1 + \Theta_{of}} - \zeta^2 \right) : \mu_2 = \zeta\omega_n \\ \omega_3^2 &= \omega_2^2 : \mu_3 = \mu_2 \end{aligned} \quad (19)$$

Through neglect of the third term indicated above, the resulting expressions for  $\omega$  and  $\mu$  from eqn (19) are functions of  $K, \zeta, \omega_n$  and  $\Theta_{of}$ , but not also of  $t$  as initially assumed.

#### Correlation with Roots-surface

From a comparison of eqns (11) and (19), the equivalent approximate closed-loop poles for large transients are identical with the small-perturbation roots evaluated at the final value  $\lambda\theta_{of}$ .

The accuracy of the approximation can be seen with reference to Figure 14. This shows three step responses of  $\lambda\theta_0$ , obtained from an analogue computer, for the process with  $K=1, \zeta=0.707$ ,

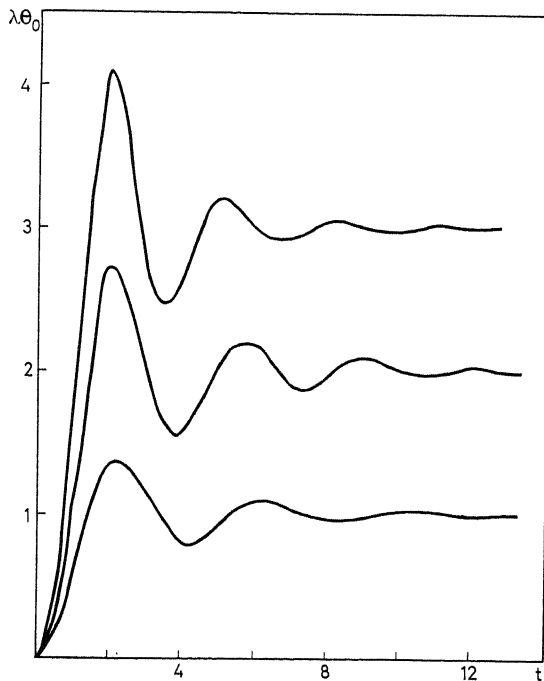


Figure 14. Analogue solutions for  $K = 1, \zeta = 0.707, \omega_n = 1$

$\omega_n = 1$ . Eqn (19) gives that the equivalent damping ratio is constant, while the equivalent damped frequency depends on  $\lambda\theta_{of}$ . For the solutions of Figure 14, eqn (19) gives values for  $\omega/\omega_n$  of 1.41, 1.68, and 1.93, while the actual values from the figure are almost constant at 1.59, 1.85, and 2.06 respectively: eqn (19) gives a constant value for  $\mu$  of 0.707 compared with the computer values of 0.48, 0.56, and 0.62 (determined from the initial overshoot), all of which tend to 0.707 as the oscillations are damped out.

#### Conclusions

The large step response of a controlled process whose small-perturbation transfer function consists of an output-dependent gain and a simple fixed pole is similar to that for a first-order linear system. The magnitude of the steady-state response depends in a non-linear way on the magnitude of the disturbance. For the transient response, time constants of nearly equivalent linear exponential form can be found in terms of the initial and final small-perturbation time constants; these describe the response with adequate accuracy for engineering purposes.

The stability of response of a controlled process whose small-perturbation transfer function consists of an output-dependent gain and a complex fixed pole-pair depends on the magnitude and direction of the disturbance and on the initial equilibrium. By Liapunov's Direct Method satisfactory general criteria have been found and in a particular case compared with a complete analysis by the phase-plane method. The 'roots-surface' associated with the system is incapable of predicting the bounds on the disturbance for stable response, which is an important restriction on the 'roots-surface' method of analysis (cf. M'Pherson<sup>6</sup>).

For a disturbance within the bounds of stability the large step response is similar to that for a second-order linear system. The damping and frequency of nearly equivalent second-order response can be adequately expressed in terms of the features of the roots surface, as is verified in some particular cases by means of an analogue computer and by an approximate method of analysis involving time-varying amplitude and phase.

*The interest and support of the UKAEA (Winfrith), and in particular of P. K. M'Pherson, Head of Dynamics Group, in connection with the project of which this work forms part is acknowledged with appreciation.*

#### References

- NECHLEBA, F. Extension of the concept of time constant. *Elektrotechn. Z.* 74 (1953) 98
- MORGAN, P. G. Definition of an equivalent time constant. *Control Data Sheet No. 24*, October 1961
- LA SALLE, J. and LEFSCHETZ, S. *Stability by Liapunov's Direct Method with Applications*. 1961. New York; Academic Press
- MURPHY, G. M. *Ordinary Differential Equations and their Solutions*. p. 168. 1960. New York; Van Nostrand
- GRENESTED, P. E. W. The frequency response analysis of non-linear systems. *Instn elect. Engrs, Lond. Monogr.* No. 126 (1955)
- M'PHERSON, P. K. Application of complex plane methods to system design. *Trans. Soc. Instrum. Tech.* 14 (2) (1962) 81

## DISCUSSION

## Authors' Opening Remarks

Since our paper was submitted further Liapunov functions have been discovered which represent improvements on the functions  $V_1$  and  $V_2$  described in the paper and illustrated by the regions  $\Omega_1$  and  $\Omega_2$  in Figure 13.

The restriction on the validity of  $V_2$  to the range  $0 < N \leq 14.25$  limits its usefulness, whilst  $V_1$  yields an over-conservative estimate of the region of stability. Using for greater convenience the following pair of equations in  $\phi_3$  and  $\phi_4$ , equivalent to eqns (18)

$$\dot{\phi}_3 = \omega_n(\sqrt{R}\phi_4 - 2\xi\phi_3) \quad \dot{\phi}_4 = -\omega_n\sqrt{R}\phi_3(1 + \phi_3)$$

where

$$\phi_3 = K\Theta'_1/R \quad \text{and} \quad \phi_4 = K(\Theta'_2/\omega_n\sqrt{R} + 2\Theta'_1/\sqrt{N})/R$$

the function  $V_3$  has been found by the method of Zubov<sup>1</sup> and the functions  $V_4$  and  $V_5$  by the same method of undetermined coefficients as used for  $V_2$ :

$$V_3 = (\phi_3^2 + \phi_4^2)/2\xi\omega_n \quad V_4 = \phi_3^2 + \phi_4^2 + 2\phi_3^3/3$$

$$\dot{V}_3 = -2\phi_3^2(1 + \sqrt{N}\phi_4/2) \quad \dot{V}_4 = -4\xi\omega_n\phi_3^2(1 + \phi_3)$$

$$V_5 = \phi_3^2 + \phi_4^2 + [2(N+2) - 4\sqrt{N+1}]\phi_3^3/3N$$

$$\dot{V}_5 = -\xi\omega_n\phi_3^2[4 + (4(N+2) - 8\sqrt{N+1})\phi_3/3N \\ + 4(\sqrt{N+1} - 1)\phi_4/\sqrt{N}]$$

In the selection of the forms for  $V_4$  and  $V_5$ , the principal object has been to produce infinite half-planes of negative  $\dot{V}$  instead of a bounded region as with  $V_2$ ; at the same time, since explicit expressions for the extent of the derived region of stability are desirable, these forms have had to be of low order. The existence of a useful region of stability from  $V_4$  has been proved by reasoning similar to that used for  $V_2$ ; the existence of the region of stability from  $V_5$  has been deduced from a theorem on Liapunov functions (Reference 3 of the paper, Theorem VI).

Using as a convenient measure of the extent of the region the coordinate  $L_n$ , where the  $\Omega_n$  contour intersects the positive  $\phi_1$  axis, the expressions for  $L_1$ ,  $L_3$ ,  $L_4$  and  $L_5$  are

$$L_1 = 0.5; \quad L_3 = 2/\sqrt{N+4};$$

$$L_4 = [\sqrt{3(N+12)(3N+4)} - (N+12)]/4N$$

$$L_5 = [\sqrt{3(N+8+4\sqrt{N+1})(3N+8-4\sqrt{N+1})} \\ - N - 8 - 4\sqrt{N+1}]/[4(N+2) - 8\sqrt{N+1}]$$

As regards  $L_2$ , it is shown in the section entitled 'Less Restrictive Bounds' that for  $N < 8$  the value of  $L_2$  is 1.0, while its value for  $8 < N \leq 14.25$  is implied in the last paragraph of that section. Figure A shows the way in which these further functions remove the restriction on the value of  $N$  required by  $V_2$ , and successively extend the zone for which stability is assured. Figure B shows the zones  $\Omega_3$ ,  $\Omega_4$ , and  $\Omega_5$  corresponding to these new functions in a particular case.

Finally, it must now be said that the Introduction reveals an underestimation of the difficulty of achieving general design techniques for a non-linear system, and it would be more appropriate to replace the phrase 'design procedures ... are being developed' (in the last paragraph), by 'design procedures ... are being sought'. However, it is believed that this study throws valuable light on the behaviour of two

output dependent processes, and as a result good progress is being made in the study of other processes of the class.

## Reference

- <sup>1</sup> ZUBOV, V. I. Mathematical methods of investigating automatic regulating systems (Leningrad, 1959), *USAEC transl. AEC-tr-4494* (September 1961) 64-68

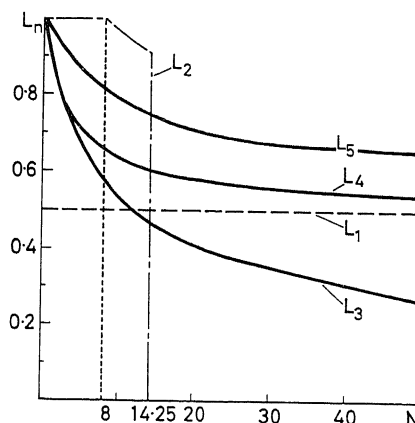


Figure A.  $L_n$  as a function of  $N$ , derived from  $V_n$ , showing a measure of the extent of the regions of stability derived from different Liapunov functions

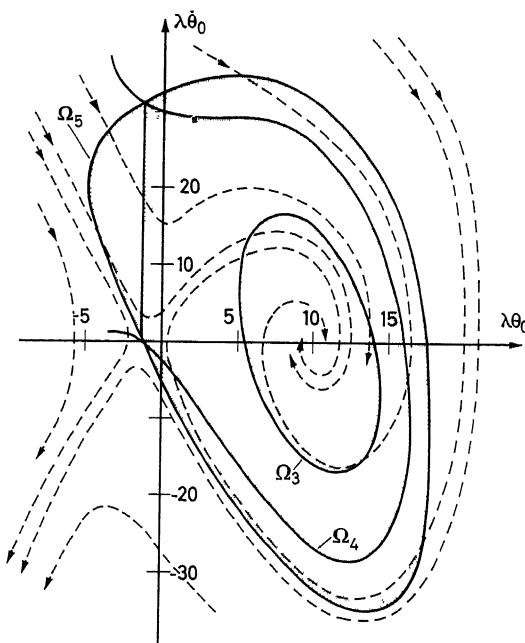


Figure B. Phase portrait of controlled process with  $\xi = 0.707$ ,  $\omega_n = 1$ ,  $K = 1$ ,  $\lambda\theta_{of} = 10$ , and  $N = 22.16$ , with regions of stability from  $V_3$ ,  $V_4$ , and  $V_5$

P. K. M'PHERSON, C. & I. Division, U.K.A.E.A., Winfrith, Dorset, England

In commenting on this paper, I would like to relate this and similar work to the needs of the control scientist who is working in the field of process control. My own field is in nuclear reactor technology and the examples I take must come from there.

This session is called 'General Problems' and, with little exaggeration, the most general problem as seen in the field sometimes seems to be to find a part of control theory that is both accessible and applicable. For example:

(1) As an exercise an attempt was made to design an optimum bang-bang controller for a large nuclear power reactor. I allowed no approximation or simplification to be made in the representation of the plant—it was to be a real engineering exercise to see if phase plane analysis would work. It did not. Not only was the problem intractable analytically (we could not reduce the problem to less than 12 dimensions) but we doubted if any existing digital computer could handle adequately a numerical treatment of the analysis.

(2) The present paper before us, where we have heard that the 'roots-surface', or a piecewise-linear approach, has failed to provide useful data on stability in the large for anything but the simplest cases.

Some people might say at this point that the problem has been solved using Liapunov techniques. I agree that Liapunov functions have been found for what the plan dynamicist would call trivial cases. In general these exercises start with the following equations to represent a nuclear reactor:

$$\dot{n} = \left( \frac{k - \beta}{l} \right) n + \sum \lambda_i c_i + S$$

$$\dot{c}_i = \frac{\beta_i n}{l} - \lambda_i c_i$$

$$k = k_t - \sum r_j T_j$$

$$\dot{T}_j = a_j n - g_j T_j$$

$n$  neutron density

$c_i$  concentration of  $i$ th group of delayed neutron emitters

$S$  neutron source

$\beta_i$  yield of delayed neutrons of  $i$ th group

$\beta = \sum \beta_i$

$\lambda_i$  decay constant of  $i$ th group of delayed neutron emitters

$k$  reactivity

$l$  neutron generation time

$T_j$  temperature of  $j$ th region in the core

$r_j, a_j, g_j$  constants

which show stability in the large for the show-down and all-power cases. For example that due to Lipkin, and Ergin

$$V_1 = n + \sum c_i + \frac{r_j}{2la_j} T_j^2$$

$$\dot{V}_1 < 0 \quad \text{if} \quad \begin{cases} r_j a_j > 0 \\ h > 0 \\ k_j \leq 0 \end{cases}$$

Letov's and Lurie's methods may be used to find other Liapunov functions. Smets has studied nuclear reactor stability using topological methods, describing functions and straightforward analysis, as well as Liapunov. He has also considered some special types of non-linear feedback. But for the more complicated cases delayed neutrons are usually omitted which is not of much use to the control engineer who relies on these delayed neutrons to make the reactor controllable.

What I wish to show is that, admirable though this work is, the systems studied cannot be imagined to represent a real nuclear reactor plant with its many interactions and non-linearities. Figure A shows a block diagram of a real nuclear reactor, being representative of the dynamics of a boiling water reactor.

It is full of non-linearities (particularly the boiling channel), transfer delays and feedback which dominate the dynamic behaviour characterized by the previous equations which, in fact, only represent the top loop round neutron production.

One would like to have a method of analysis which would indicate comparatively quickly the stability in the large of these multivariable, multiloop, non-linear systems without having to spend considerable effort on analogue or numerical analysis. The only theory that is accessible to this class of problem at the moment is the much despised linear theory. So it seems reasonable to ask Professor Maclellan and Mr. Kerr to devise a method which would give us information on stability in the large from a consideration of the type of linear models used in system synthesis. Although they have given us some very interesting results, and all of us concerned with the project have acquired a considerably deeper insight into non-linear stability problems, one must conclude that this approach is unlikely to yield the results that had been hoped for. So the field is still wide open, and one begins to wonder which approach might be better: a full-scale attack on Liapunov to find the rules with which to provide Liapunov functions for this class of multiplicative output dependent system, or resort to numerical analysis any time an analytical difficulty is encountered. I would like to end by asking the following intentionally provocative question: in view of the fact that numerical analysis on digital computers can reduce the most complex engineering problems to a tractable form, is it worth devoting any significant research effort to the development of analytical methods of solution which at the best are only likely to provide qualitative answers?

Liapunov functions are known for the impulsive input case

$$k(t) = h\delta(t) + \sum r_j a_j e^{g_j t}$$

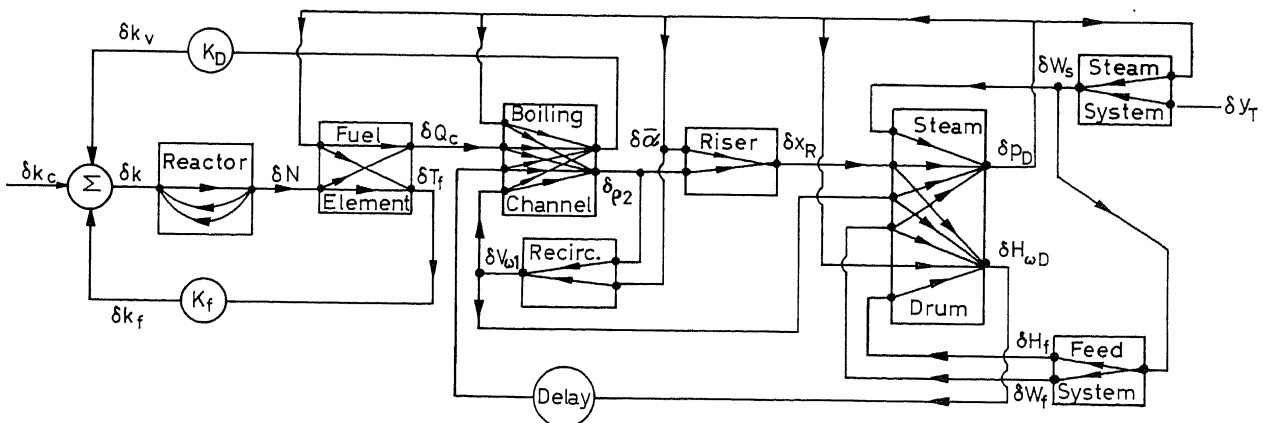


Figure A. Simplified flow chart for boiling water reactor plant

C. N. KERR and G. D. S. MACLELLAN, *in reply*

We noted with interest that T. J. Williams also drew attention to this general problem in his Congress Survey Paper, in the sections entitled 'The Uses of Process Dynamics in Process Plant Studies' and 'The Formulation and Use of Theoretical Models'. We agree with Mr. M'Pherson that the locally linearized approach to analysing systems of the output-dependent class has so far proved to be of little value except in the simplest cases. It is appropriate here to point out that the small-signal, output-dependent transfer function of any process may not be unique to that process, and in consequence the method inevitably has its limitations. For example, the following transfer function

$$\frac{\delta\theta_o(p)}{\delta\theta(p)} = \theta_{oe} \frac{p + a_{00}}{p(p + a_1)}$$

and the corresponding roots-surface belong to both the process in eqn (2) of the paper and the process described by the equation

$$\frac{d^2\theta_o}{dt^2} + a_1 \frac{d\theta_o}{dt} - \left( a_{00}\theta + \frac{d\theta}{dt} \right) \theta_o = 0$$

The large-scale transient responses of these different controlled processes will obviously differ, yet any deductions from the features of the roots-surface would apply to each.

Since a general locally linearized approach to these systems appears to be unfruitful, Mr. M'Pherson poses his challenging question. In our opinion, it is still desirable to employ whatever analytical methods are valid and useful to achieve general results in analytical form, even though such methods may be more complex than the existing linear methods. We feel that the effort expended in utilization of Liapunov's method, for example, pays off finally by providing definite quantitative information about the behaviour of systems of this class, as is demonstrated by the case of the second-order controlled process of the paper.

C. N. SHEN, *Mechanical Engineering Department, Rensselaer Polytechnic Institute, Troy, New York, U.S.A.*

The authors of this paper have presented lucid examples of control theory techniques. An output equilibrium level  $\theta_{oe}$  can be expressed in terms of an input equilibrium level  $\theta_e$  and vice versa. It seems to

me that in eqn (4) of the paper the use of eqn (3) to eliminate  $\theta_{oe}$  instead of  $\theta_e$  becomes an 'input-dependent process'. Would the authors care to discuss the word 'output' in the title of their paper?

There is no equilibrium level  $\theta_{oe}$  for the reactor kinetics in eqn (2). The neutron density for a critical reactor could take any level at equilibrium condition, thus eqn (3) should be modified. The application of the authors' method to the problem of reactor kinetics with six groups of delayed neutrons would be of great interest. In addition, it would be helpful if the authors would indicate the physical significance of their integral criterion on the accuracy of the result with reference to Figure 4.

I believe that the problems for output dependent processes are very important. Thus the purpose of the discussion is to generate interest in this field of study.

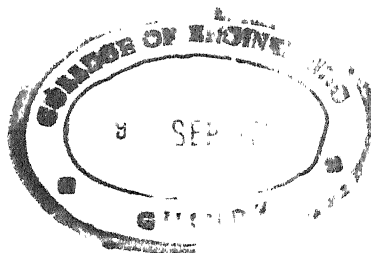
C. N. KERR and G. D. S. MACLELLAN, *in reply*

Professor Shen is correct in stating that eqn (3) may be used to eliminate  $\theta_{oe}$  rather than  $\theta_e$  from eqn (4). However, a situation may arise, as he points out, where the value of  $\theta_e$  is zero; this occurs if the transfer function has a pole at the origin. In such an equilibrium state, the output may have any value, and the transfer function depends only on  $\theta_{oe}$ . Furthermore, for the two processes considered the block diagrams of Figures 1 and 5 show how the instantaneous gain can be regarded as directly dependent on the output. For these reasons, the general class of process is regarded as output, rather than input, dependent.

There is no need to modify eqn (3) for the case of a nuclear reactor in which the coefficients  $a_0$ ,  $b$  and  $b_0$  are zero. At equilibrium, either  $\theta_e$  or  $\theta_{oe}$  (or possibly both) must be zero, and the only non-trivial situation is that where the equilibrium neutron flux is non-zero.

We agree that an investigation of the reactor with all six groups of delayed neutrons would be of considerable interest. However, it has been found helpful to start this general investigation with processes of lower order, and it is intended to extend the investigation successively to higher orders.

The time integral of the transient deviations of the output from its final value provides a convenient measure of the quality of the response. By equating the value of this integral arising from the approximate solution to the value from the exact solution, a 'best' approximation is obtained.



# REPORTS

---

B. N. PETROV\*

Automatic control theory is a new field which in recent years has been making exceptionally rapid progress and has undergone radical qualitative changes.

We have witnessed fundamental developments in the theory of non-linear system stability and in the theory of linear and non-linear sample-data systems. New design approaches have been suggested for self-adjusting and self-adaptive systems, constituting a new type of automatic control. The interest of the specialists has been focused on optimal control. Various techniques for the design of optimal systems based on different performance criteria have been developed, notably, Pontryagin's maximum principles and Bellman's dynamic programming.

Stochastic methods have found their way into practically every field of automatic control theory. More attention has been devoted to structural problems and to hybrid and invariant systems; the sensitivity problem has been formulated; new results have been obtained in finite automata and switching circuits theory.

Principles of self-organization and learning machines, problems of pattern recognition and the 'man-machine' problems have lately been extensively investigated. These as well as bionics and the theory of large-scale systems are purely cybernetic matters, which broadened the scope of modern automatic control.

Remarkable survey papers on fundamental problems of automatic control which cite most important recent results in our field of art, have been a very valuable contribution to the scientific programme of the Congress.

The Second I.F.A.C. Congress has clearly demonstrated these qualitative changes and marked progress in automatic control theory. Moreover, it is equally important to stress the fact that this meeting has provided an opportunity for the scientists from all parts of the world to discuss in friendly atmosphere the problems they are working on.

In the brief summary, which I have the privilege to deliver to the Congress, it is not possible to refer to the authors personally and to evaluate new results which have been reported. Therefore I offer my apologies to the speakers and will concentrate only on main issues, being unable to cover all branches of automatic control theory presented here.

The Congress was high-lighted by the concepts of optimal control theory. Optimal control has made essential progress and optimal control technique has been generalized for many

types of automatic control systems (such as, sample-data, distributed parameters, time-lag and digital systems, not to mention self-adaptive control where optimal and sub-optimal processes are inherent).

Control algorithms based on various performance criteria for different plants, including industrial processes and utility systems, have been cited in many papers and digital computers are widely used for practical implementation of these algorithms.

The stability problems have been solved for a number of new types of control systems, and particularly for self-adaptive systems. Liapunov's methods, created more than 70 years ago, had a spectacular revival with the advent of these new types of systems and with the progress in computational procedures and equipment. It has been shown that optimal controllers designed on this basis and utilizing dynamic programming or maximum principle are inherently stable.

These results are obtained in a number of papers devoted to the control system synthesis and analytical design.

Although the Congress has not been marked by essentially new methods of statistical control system theory, the papers and their discussions have demonstrated ever-spreading applications of stochastic approaches in the control system analyses and synthesis, and in the evaluation of optimal control laws. A new theory of dual control has been developed which makes use of decision function theory.

It is extremely interesting to note that many papers reflect further research of the problems reported at the First I.F.A.C. Congress and solve a number of problems which were the topic of discussions in Moscow.

Sample-data and digital systems have received great attention at this Congress. Papers dealing with foundations of non-linear sample-data theory show the way to investigate stability of the systems to evaluate their performance and periodical oscillations when different types of non-linearities are present.

More effort has been directed to the solution of structural problems and to the search for new principles of automatic control, combining high accuracy with low sensitivity to parameter variations and invariance to external disturbances.

Here we have witnessed new results associated with optimal control problems. New important features and efficiency of variable structure systems have been investigated, and design technique for invariant sample-data and hybrid systems developed. A new problem of configuration automatic control for complex systems has been stated and a way to its solution proposed.

\* Chairman of the I.F.A.C. Technical Committee on Theory.

The well-established theory of linear systems has made new progress. I mean the solution of the quadratic, estimates inversion problem.

Papers devoted to self-adaptive systems reflect essential progress in the state of art. Professor Truxal, who expressed his disappointment in the absence of new results in this field during the past six years, has every reason to be satisfied. I am referring to the papers in which Liapunov's methods, describing function technique and other approaches, were used to evaluate self-adaptive system stability, the way to speed up an optimum mode of operation was proposed, the fundamentals of identification theory were investigated and the number of other effective techniques were developed.

Even more papers deal with analysis and design for particular types of self-adaptive systems.

Many participants at the Congress were involved in the discussion of learning machines and systems, this being practically the first opportunity to consider these problems at large.

A number of interesting papers dealt with new cybernetic lines of thought in automatic control theory as well as to the 'man-machine' problem. These papers contain the basic theory with a number of specific block-diagrams and examples of learning machines which will find a wide application in future.

It is to be regretted that at the Congress problems of bionics have not received the attention they deserved.

In spite of the fact that the symposium on relay systems and finite automata was held less than a year ago, the papers in this field presented some new concepts. These papers considered once more that the theory of finite automata is being incorporated in the general automata control theory.

As a large river taking numerous tributaries brings its waters to the sea, so the general theory of automatic control, forming from the early independently-developed branches of technical sciences and enriching constantly with new methods, gives all its power and possibilities to the technical progress to the better future of mankind.

## J. H. WESTCOTT\*

It seems to me that we are at present in a period of consolidation of theory, and this does not lead to quite such a spectacular show of sparks as in the period before it, when new ideas are flashing around everywhere. About 5 or 6 years ago we reached the period in which the river of our endeavour in theory had become very thoroughly frozen, although we were not particularly aware of it. Perhaps we were rather satisfied with the easy way in which we could skate over the difficulties that lay below, but nevertheless, underneath subtle movements were at work and these movements began to generate a growing amount of heat as discussion broadened, until about the time of the founding of IFAC, the whole river began to break up into icefloes. I regret to say that some of our unfortunate colleagues were unceremoniously ducked into the water and some, it is rumoured, were carried out to sea never to be heard of again. This was a time of great fluidity; old theories were brought out again and dusted; new theories were being launched every month; the chase was on and we were sometimes a little intoxicated and a little incautious. By the time of the first Congress in Moscow the river was flowing in a new direction. From all countries there were scientists and engineers who often independently were working with the new ideas of extremal control, whether in the form of actual equipment, simulated systems or in the currency of the mathematical ideas themselves. Each group had its own terminology, its own preferences, its own method of framing the problem. In the river there was still one or two icebergs floating about and nobody knew for sure whether they were 'band-waggon-for-getting-on', iconoclasts of destruction, or just a bad dream. They have now done their work of sweeping away some of the excrescences and have for the most part sailed out to sea leaving a certain amount of untidiness behind them.

In the three years since then we have seen a good deal more order return to the river; some traffic is beginning to flow towards the market place. There are still many pockets of undeveloped

territory, there are even one or two misconstructions, which may have to be dismantled; but by and large the immediate impasse has been breached and we can none of us complain of having nothing to do. Let me take you on a brief tour of the river as I now see it, and since we are in Basle we shall make our crossing by one of those ingenious and picturesque ferry boats which have no engine or source of power but rely entirely on the skill of the steerman to exploit the prevailing currents.

The main current of progress of the Congress, in theory, appears to have two branches:

- (a) uses of the variational principle;
- (b) uses of statistical estimation theory.

The first gives us optimal theory; in the noise-free case using trajectory methods. The methods of characteristics enables us to split the equations into a set of dynamic state equations and an adjoint state of co-state equations. Thus we have a two-point boundary value problem.

The real difficulty lies in the numerical solution of problems of realistic size and complexity. There is much stumbling here and I sense that little shows for the prodigious effort that is, and needs to be, made to establish method. It is possible that a full-scale attack by specialist numerical analysts would save much unrequited effort here, although there are reasons to doubt whether this would be successful without the incentive of engineering motivation. So perhaps we shall have to struggle on our own. The real target here is not just a numerical solution, but a genuine synthesis method, preferably realized in feedback form. I think by the next Congress we will have such a method.

The second branch deals with control in the presence of random disturbances; here the remedy of the method of characteristics is denied us and the minimizing partial differential equation must be solved directly. The Chapman-Kolmogorov equation shows how the random effects are diffused through the system modified by its dynamic characteristics. Our numerical difficulties, even under the simplification implied by the Fokker-

\* Vice-Chairman of the I.F.A.C. Technical Committee on Theory.

Planck equation, are now increased. The handling of the stochastic case (that is under conditions of statistical equilibrium), is near to the practical limit to achievement here.

Finally, to take the problem of greatest practical interest, and hence, it always seems, maximum mathematical difficulty, we come to the fully adaptive problem. Here we have to introduce the idea of statistical estimation of possibly-varying parameters. The identification problem is one aspect of this and we have had some very practical contributions on this subject. In making control manoeuvres in an adaptive system we have a double motive:

- (a) to improve estimation of parameter values;
- (b) to improve control using current parameter values.

The compromise has to be made using statistical decision theory, and we have received helpful suggestions for this. I would like to make a personal comment on theoretical developments. There are three possible ways of exploiting mathematical theory:

- (1) to make use of existing theory;
- (2) to re-invent old theory;
- (3) to invent new theory.

I put them in order of frequency of occurrence. Now in using variational methods we are clearly using established theory, but as we have seen when wishing to become numerate, we have had difficulties. In struggling against these difficulties we have often re-invented old theory. This is no reflection on the quality of this work, for it was no more easy to develop for a second time; but we have been in this numerical difficulty before and

much valuable work lies unremembered in the old works on the numerical solution of differential equations. It could well be worth while to go through the vaults before embarking on the prodigious labour of outright invention.

I have talked about the papers presented to the Congress. There is another facet of the Congress which is perhaps more important to workers in theory than in other spheres since they deal in a currency of ideas and that is the question of making personal contacts with other theorists. I have found this aspect of the Congress most valuable, but I cannot disguise from myself that it is still a difficult and time-consuming activity finding the man one wants to talk to. Between sessions a sort of Brownian motion sets in. As a distinguished colleague said, using what Professor Gerecke called the international language of IFAC, 'all people are chasing each other trying to meet. It means, although I should not say it, half the time you are talking to the wrong people'. This is still one of the unsolved problems of congresses.

Professor Truxal has let you into the secret of these light signals on the speakers' rostrum. Fortunately for me they have not had time to fit the new one that he suggested should be added. During the first session of the Congress, when I presented my paper, the secret of how these lights worked was such a well kept one that even the chairman did not know it. In his desperation at my loquacity, he crept up behind me without my noticing, as I was in full spate, and as I paused for breath said 'Quit stalling, and give us the new stuff'. I can think of no better slogan for the next Congress to speed the parting guests at the end of this Congress, than 'quit stalling and give us the new stuff'.

